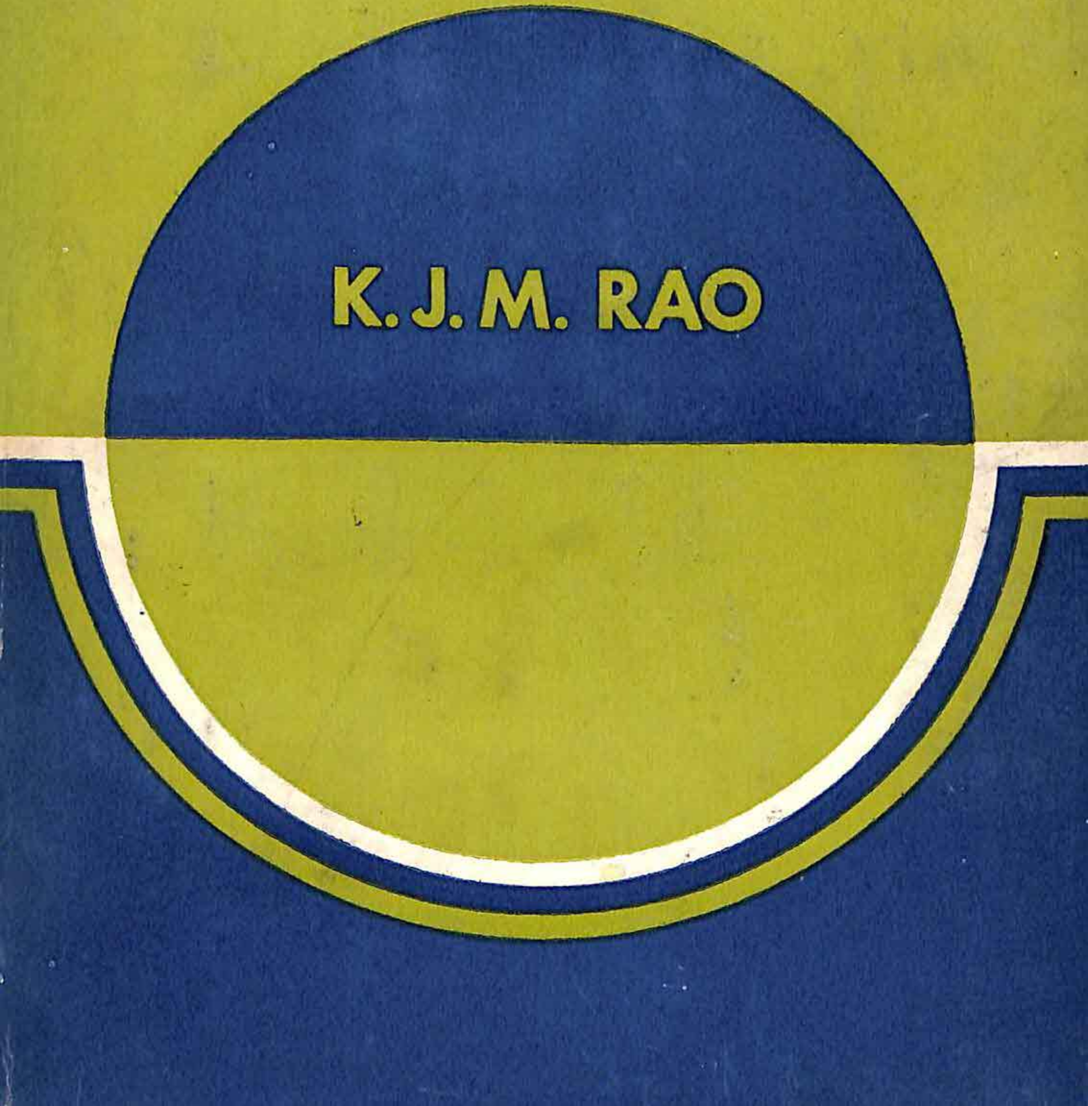
INTRODUCTION
TO
SEMICONDUCTOR
DEVICES



K. J. M. RAO

D1437/4

రూ 42/-

INTRODUCTION TO SEMICONDUCTOR DEVICES

(A companion volume to author's other book entitled "Introduction to Electronics")

K.J.M. RAO
National Defence Academy
Kharakvasla, Pune

S. C. E. R. T., W.B.
Library
Calcutta
25 3,

OXFORD & IBH PUBLISHING CO. PVT. LTD.
New Delhi Bombay Calcutta

S.C.E.R.T. W.B. LIBRARY
Date 23.12.98
Accn No 9787

© *K.J.M. Rao, 1977*
Second printing 1987

ISBN 81-204-0211-1

Published by Mohan Primlani for Oxford & IBH Publishing Co. Pvt. Ltd.
66 Janpath, New Delhi 110 001 and printed at
Chaman Offset Printers, New Delhi 110002

2-F7-11

Foreword

In recent times the subject of electronics has become quite important not only to engineers but also to physicists both at the undergraduate and postgraduate level. An understanding of the elements of electronics and their applications has become necessary now in all the fields of science and technology. The need to understand the basic physical principles of transistors and the other more recently discovered new devices is felt increasingly by the electronic engineers. In view of this, a deeper understanding of the physics of transistor devices is necessary for exploiting the full capabilities of these devices for development of electronic instruments. There is also a dearth of books on this subject by Indian authors.

The present volume is expected to fill this gap by providing useful introduction to the study of semiconductor devices. This book covers the field of physics of semiconductors fairly comprehensively and covers the essential theories which led to an understanding of the physics of semiconductors including the latest semiconductor devices. Shri Rao has not only wide teaching experience in the subject to be able to present the material in a very lucid manner but has also the necessary research experience to see that the material that is presented is up-to-date. I have no doubt that this volume will be useful to all physicists and engineers who wish to have a fair understanding of the subject and who wish to pursue research work in this important area. I do hope that this volume will be as received as his earlier volume entitled "Introduction to Electronics".

B. Ramachandra Rao

Preface

The book is written as a companion volume to the author's earlier book 'Introduction to Electronics'. In this age of ever expanding scientific horizons the topic semiconductor devices has assumed great importance, particularly for students of Physics and Engineering. It has been included in the syllabus by a number of universities in India.

The book is based on the lectures delivered by the author to the students of Poona University. Switching devices which were covered in the earlier book 'Introduction to Electronics' have not been included in this book.

The need for a sound theoretical background, which is essential to the proper understanding of the devices, cannot be overemphasized. However, the reluctance of the manufacturers to make available the exact data, has prevented the author from giving numerical values.

The author is extremely grateful to Prof. B.R. Rao, Vice Chairman, U.G.C. for his valuable guidance, useful suggestions and encouragement and also for writing the foreword. Thanks are due to Prof. M.R. Bhiday, Head of the Physics Department, Poona University, whose encouragement, advice and guidance were of immense help in writing this book.

The author is grateful to Dr. H.K. Srivastava, Principal, National Defence Academy and Prof. S.S.P. Bhatnagar, Head of the Physics Department, N.D.A. for their encouragement. He is also grateful to Professors G.K. Bhagawat of I.I.T., Bombay,

J. Das of I.I.T., Kharagpur and B.V.T. Rao of I.A.T., Pune for their valuable suggestions.

This book is written under the University Level Textbook Writing Programme of the University Grants Commission and the author is grateful to the Commission for providing financial assistance for the preparation of the manuscript. He wishes to express his sense of gratitude to the authorities of the N.D.A. and the Defence Ministry for kindly permitting him to undertake the job.

He is also grateful to Mr. M.S. Sejwal and M/s Oxford & IBH Publishing Co. for the cooperation extended in bringing out the book in a short time.

The author will gratefully welcome any critical suggestions for improving the quality and contents of this book from all users.

K.J.M. RAO

Contents

Foreword

Preface

1. **Introduction**
   Schrodinger's equation and crystal structure; Semiconductors; Effective mass; Carrier densities; Variation of mobility with field strength—hot electron effect; Diffusion; Carrier generation and recombination; Continuity equation; Carrier density profile—diffusion length

2. **Junctions** 26
   Fabrication of junctions; Band diagram of a $P-N$ crystal—barrier voltage; Depletion layer; Graded junctions; Current voltage relation—Shockley equation; Forward bias; Quasi-Fermi levels; Temperature dependence of reverse saturation current; Carrier generation and recombination within the depletion layer; Junction breakdown; Charge storage—Admittance of a junction; Transient behaviour of a diode

3. **Transistor** 57
   Current flow through a transistor; Current calculations; D.C. characteristics; Common base configuration; Common emitter configuration; Voltage limitations;

A.C. behaviour of a junction transistor; Admittance of a forward biased junction; y and h parameters of a transistor; Equivalent circuit of a transistor; Frequency dependence of β; High current effects in a transistor; Thermal effects; Large signal theory; Transistor as a switch; Graded base transistor; Punch through voltage

4. **Junction Field Effect Transistor** 119

Structure of a field effect transistor; Characteristics of a FET; Transconductance; Non-uniform impurity distribution in the channel; Series resistances; Equivalent circuit of FET

5. **MOSFET** 132

MOS structure; Departures from ideal; Effect of charges; Principle of operation of MOSFET; Current voltage relation; Applications of MOSFET

6. **Microwave Semiconductor Devices** 153

Parametric amplification; Manley-Rowe equations; Varactor; Snap back diode: PIN diode; NπP diode; IMPATT—Read diode; Tunnel diode; RWH mechanism —Gunn effect; Dielectric relaxation time; Modes of operation

7. **Opto Electronic Devices** 193

Photoconductivity; Photoconductors; Photodiodes; Solar cell; Photodiode detectors; Luminescence; GaAs diode lamp; Spontaneous and stimulated emission; Two-level system; Injection laser

Index 219

1. Introduction

In 'Introduction to Electronics' physics of the semiconductor materials and current flow in semiconductors were discussed in an elementary manner. The discussions and results obtained were sufficient to understand the functioning of junction diodes and transistors. There are a number of semiconductor devices other than the junction diodes and transistors whose functions can only be explained on the basis of quantum mechanics. In this chapter physics of semiconducting materials and current flow are discussed. The discussions are restricted to the results essential to understand the functioning of the devices and no attempt is made to derive the results.

1.1.1 *Schrodinger's equation and crystal structure*

It is experimentally found that a particle having a momentum p behaves as a wave of wavelength h/p and a wave of energy E behaves as a particle of momentum E/c. The electrons moving round the nucleus of an atom therefore have wave properties associated with them. Bohr's postulates of atomic theory are found to be the consequences of the wave nature of the electron. The electron motion is governed by a wave equation first given by Schrodinger and is called the Schrodinger's equation. The Schrodinger's equation is

$$\frac{-\hbar^2}{2m}\nabla^2\psi + V\psi = j\hbar\frac{\partial\psi}{\partial t} \qquad (1.1)$$

where m is the mass of the electron, $\hbar=h/2\pi$, V the potential energy and ψ is called the wave function. When the above equation is solved for ψ we obtain a wave equation governing electron motion. The real part of $|\psi^2|$ at any point gives the probability of locating the electron at the point and is referred to as the probability amplitude. From the solution for ψ the energy and momentum of the electron are found.

1.2.1 *Crystal structure*

In a crystal atoms occupy well defined positions, the lines joining the centres of atoms gives a geometrical structure called the lattice. The energy and momentum of electrons in a crystal are found by formulating the Schrodinger's equation for electrons in the crystal and solving it.

Instead of considering a three dimensional lattice and formulating the Schrodinger's equation it is convenient to solve the Schrodinger's equation for a one dimensional case and extend the results to the real situation in which the lattice extends in the three dimensions. For the simple one dimensional case we find that atoms are arranged in a row. When an electron is situated near an atom it is under the influence of the electrostatic forces due to the nucleus and the extra nuclear electrons, as such it has a negative potential energy. As the atoms are arranged in a row the electron is located in a region of equidistant potential wells. The potential energy V in the Schrodinger's equation is therefore not a constant but a periodic function with a period equal to the spacing between atoms, a. The probability amplitude or the wave function is given by the expression

$$\psi=\mu\,(k.x)\exp j\left(k_x-\frac{E}{\hbar}\,t\right) \tag{1.2}$$

where $\mu\,(k\cdot x)$ is a periodic function in k and x and is called the Bloch function. The exponential part of expression (1.2) represents a travelling wave along the X-axis. k_x which is identified as a wave number in a wave equation is called the wave vector as $k=\frac{2\pi}{\lambda}=\frac{2\pi p}{h}=\frac{p}{\hbar}$. k_x also represents the momentum along the X-axis and is called the crystal momentum, it is the momentum of phonons or the lattice vibrations. E is the energy of the electron. The expression is similar to that of a modulated wave of wave number k and angular frequency $\omega=E/\hbar$. The modulation represents the presence of the electron in the crystal, therefore the group velocity of the wave is the particle velocity and can be found as

$$v=\frac{\partial\omega}{\partial k}=\frac{1}{\hbar}\cdot\frac{\partial E}{\partial k} \tag{1.3}$$

The velocity of an electron is thus a function of its energy and wave vector k.

The energy of the electron in a crystal is given by the expression

$$E=E_0-2A_x \cos k_x a \tag{1.4}$$

where E_0 and A_x are constants and k_x is the value of the wave vector along the X-axis. From the expression we find that the electron in a crystal can have any energy within the limits E_0-2A_x and E_0+2A_x. Thus the allowed energies for electrons in crystals are spread into bands. In a band the energy increases continuously with increase in k as shown in the figure.

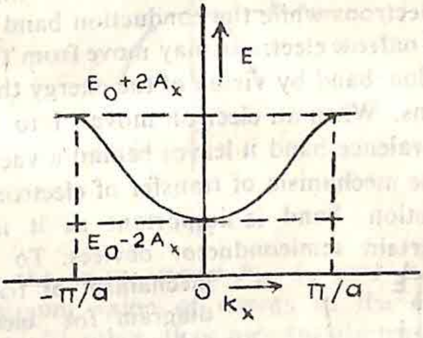


Fig. 1.1.

All allowed energies for an electron in a band are confined to a region in the momentum space bounded by the limits $-\pi/a$ and π/a. The limits $-\pi/a$ and π/a are referred to as the edges of the Brillouin zone along the X-axis. Similarly the edges of the Brillouin zones along the Y and Z axes are π/b and π/c, where b and c are the atomic spacings along the Y and Z axes. When the atomic spacings are the same in all directions the wave vector-energy diagram is a sphere otherwise it is an ellipsoid.

There are a number of energy levels for electrons in an atom. Each energy level spreads into a band. We thus see that in the crystal there are a number of energy bands. The width of a band depends on the tightness of binding of the electron, the tighter the binding narrower is the band. Therefore the bands due to the lower energy states are narrow while the bands due to the higher energy states are wide. The bands are separated from one another by energy gaps which are referred to as forbidden gaps as there are no allowed energy levels within the energy gaps. The bands of interest are the energy band due to valence electrons called the valence band and the energy band due to the ionisation levels called the conduction band. There are a number of sub-bands in the conduction band due to the interactions of the electrons with the lattice and electron spin.

1.2.2 *Semiconductors*

The electrical properties of a material depend on the width of the forbidden gap. (In semiconducting materials the forbidden gap is about 1 eV.) At absolute zero of temperature the valence band is filled with electrons while the conduction band is empty. If the forbidden gap is narrow electrons may move from the valence band to the conduction band by virtue of the energy they acquire from thermal vibrations. When an electron moves into the conduction band from the valence band it leaves behind a vacant energy state called 'hole'. The mechanism of transfer of electrons from valence band to conduction band is important as it is the basis of operation of certain semiconductor devices. To understand the mechanism of transfer the $E-k$ diagram for electrons in the valence and conduction band are drawn. When the X, Y, Z components of the wave vector are equal the $E-k$ diagrams for electrons in conduction and valence bands are parabolas as shown in the figure. (The $E-k$ diagram for the electrons in the valence band is drawn inverted as the energy of the vacant energy state is maximum at the lower edge of the band.) As the maxi-

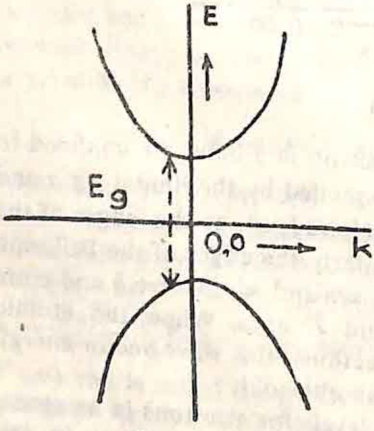


Fig. 1.2.

mum value of energy in the valence band and the minimum value of energy in the conduction band occur at the same value of k the semiconducting material is called a direct band gap semiconductor. An electron in the valence band can move into the conduction band when it acquires an energy equal to the difference in the energies for any value of k. As the momentum in the valence and conduction bands is the same direct transition between bands is possible. GaAs is one widely used direct band gap semiconductor.

The minimum value of energy in the conduction band occurs at a value of k other than zero for silicon and germanium as the

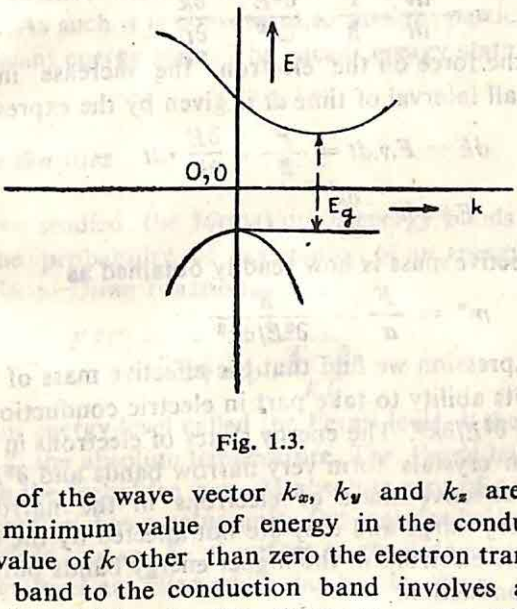


Fig. 1.3.

components of the wave vector k_x, k_y and k_z are not equal. When the minimum value of energy in the conduction band occurs at a value of k other than zero the electron transition from the valence band to the conduction band involves a change in momentum in addition to the change in energy. If the electron in the valence band is to move into the conduction band it has to acquire an additional energy equal to the band gap and also an additional momentum equal to the momentum difference as momentum is to be conserved in any transition. The transitions from the valence band to conduction band or vice versa are possible only when the necessary momentum difference is obtained from lattice vibrations. The electron can also move from one band to another in a number of steps through allowed energy states in the forbidden gap due to impurities and crystal imperfections.

1.2.3 *Effective mass*

As the electrons in a crystal are located in a field of varying potential the effective mass is different from the mass of a free electron. The effective mass of an electron is found as the ratio of the force applied to the acceleration produced by the force. For a one dimensional case we found that the velocity is related to the wave vector by the expression

$$v = \frac{1}{\hbar}\,\frac{\partial E}{\partial k} \tag{1.3}$$

The acceleration of the electron

$$a = \frac{dv}{dt} = \frac{1}{\hbar}\cdot\frac{\partial^2 E}{\partial k^2}\cdot\frac{\partial k}{\partial t} \tag{1.5}$$

If F is the force on the electron, the increase in energy dE during a small interval of time dt is given by the expression

$$dE = F.v.dt = \frac{F}{\hbar}\cdot\frac{\partial E}{\partial k}\cdot dt$$

and

$$F = \hbar\cdot\frac{\partial k}{\partial t}$$

The effective mass is now readily obtained as

$$m^* = \frac{F}{a} = \frac{\hbar^2}{\partial^2 E/\partial k^2} \tag{1.6}$$

From the expression we find that the effective mass of the electron and with it its ability to take part in electric conduction depend on the value of $\partial^2 E/\partial k^2$. The energy states of electrons in inner orbits of atoms in crystals form very narrow bands and $\partial^2 E/\partial k^2$ is very small. The effective mass of electrons in the narrow bands is therefore very large and they are not affected by the applied electric field. The electrons in the higher energy bands only take part in electric conduction.

We have noted earlier that the E—k plot for electrons in a band goes through a maximum and a minimum. The electrons at the bottom edge of the band have minimum E and $\partial^2 E/\partial k^2$ is positive while the electrons at the top edge of the band have maximum E, $\partial^2 E/\partial k^2$ is negative for electrons at the top edge of the band. As a result the electrons at the lower edge of the band have a positive effective mass and the electrons at the top edge of the band have a negative effective mass. The negative effective mass implies the movement of the electron opposite the direction of the force.

1.2.4 *Electrons and holes*

In a semiconductor, the forbidden gap separating the valence and conduction bands is narrow, of the order of 1 eV. Some electrons are excited from the valence band into the conduction band due to the thermal vibrations.

When an electron moves from the valence band it leaves behind a vacant energy state with a positive charge. Moreover as the electron moving from the valence band is located at the top edge of the band it has a negative effective mass and its absence in the valence band is associated with a positive mass. The vacant energy state in the valence band therefore has a positive charge and a positive mass. As such it is convenient to assign particle properties to the vacant energy state. The vacant energy state is called a hole.

1.2.5 *Carrier densities*

Till now we studied the formation of energy bands in a semiconductor. The probability of occupation of an energy state E is given by the Fermi Dirac function

$$F(E) = \frac{1}{1+\exp \dfrac{E-E_f}{kT}} \tag{1.7}$$

where E_f is an energy level called the Fermi level, k the Boltzman constant and T the absolute temperature. The Fermi level is generally located in the forbidden gap. At absolute zero of temperature the Fermi function has a value unity for energies below E_f and 0 for energy values above E_f, therefore at absolute zero the bands below the Fermi level are filled while the bands above the Fermi level are empty.

At temperatures above absolute zero the probability function has a non-zero value for energies above E_f indicating the existence of electrons in the bands above the Fermi level. The electrons in the valence band are excited across the band gap into the conduction band due to the energy they acquire from thermal vibrations. The densities of electrons in the conduction band and holes in the valence band can easily be calculated from a knowledge of the density of energy states in the bands and the probability of their occupation. In 'Introduction to Electronics' the densities of electrons and holes in the conduction and valence bands are calculated as

$$\left.\begin{aligned} n &= \frac{4\sqrt{2}}{h^3}\,(\pi m'\,kT)^{3/2}\cdot\exp\frac{E_f-E_c}{kT} \\ p &= \frac{4\sqrt{2}}{h^2}\,(\pi m''\,kT)^{3/2}\cdot\exp\frac{E_v-E_f}{kT} \end{aligned}\right\} \qquad (1.8)$$

m' and m'' are the effective masses of electrons and holes in the conduction and valence bands, E_c the energy corresponding to the lower edge of the conduction band and E_v the energy corresponding to the top edge of the valence band.

In an intrinsic semiconductor the densities of electrons and holes are equal. If it is assumed that the effective masses m' and m'' are equal it is clear from the above equations that the Fermi level is located at the middle of the band gap.

The product of the densities of holes and electrons in the semiconductor pn is obtained from the expression (1.8) as

$$\begin{aligned} pn &= \text{const. } T^3\ .\ \exp\frac{-(E_c-E_v)}{kT} \\ &= \text{const. } T^3\ .\ \exp\frac{-E_g}{kT} \qquad (1.9) \\ &= n_i^2 \end{aligned}$$

The product pn is a constant depending on the temperature and the width of the forbidden gap. It does not depend on the impurities introduced as long as the impurities do not change the width of the forbidden gap.

1.2.6 *Doped or extrinsic semiconductors*

When impurity atoms are introduced in a semiconductor material, energy levels are provided within the forbidden gap. A fifth group impurity atom (Al or P) introduces a filled energy state below the edge of the conduction band. The energy difference between the energy state and the lower edge of the conduction band is approximately 0.01 eV and the electron from the filled state easily moves into the conduction band leaving behind an ionised impurity atom. (The fifth group impurity atom is positively charged as it loses an electron.) Each fifth group impurity atom contributes an electron to the conduction band and the fifth group impurity atoms are called donor atoms. The material doped with fifth group impurities is called an *N*-type semiconductor.

A third group impurity atom introduces a vacant energy state above the edge of the valence band and the energy state is readily filled by an electron from the valence band as it is located very near the edge of the valence band. The movement of an electron from the valence band leaves behind a hole. Third group impurity atoms therefore introduce holes and they are called acceptor atoms. The material doped with third group impurity is referred to as a *P*-type semiconductor. The third group impurity atom in the semiconductor acquires an electron from the semiconductor material and as such is negatively charged.

We thus see that the introduction of impurities increases the density of one type of carrier or another. As the product of holes and electrons in a semiconductor is a constant depending on the width of the energy gap and temperature the introduction of impurities results in an increase in the density of one type of carrier and a reduction in the density of the other type of carrier. The carriers introduced by the impurities are called majority carriers and the other type are called minority carriers. The low value for minority carrier density is due to added recombination.

If N_d is the donor impurity density, N_a the acceptor impurity density, p the density of holes and n the density of electrons in a sample of semiconductor material we have from the condition of charge neutrality

$$N_d+p = N_a+n$$

$$n = (N_d-N_a)+p$$

$$= (N_d-N_a)+\frac{n_i^2}{n}$$

Solving the quadratic in n we have

$$n = \frac{(N_d-N_a)\pm\sqrt{(N_d-N_a)^2+4n_i^2}}{2}$$

$$\approx (N_d-N_a) \text{ when } (N_d-N_a) \gg n_i \tag{1.10}$$

The electron density in the *N*-type semiconductor equals the difference in the donor and acceptor impurity densities when they are large compared to the intrinsic density, n_i.

Similarly the hole density in a *P*-type semiconductor is given by the expression

$$p = (N_a-N_d) \tag{1.11}$$

1.2.7 *Location of the Fermi level in a doped semiconductor*

We found that the Fermi level is situated at the middle of the band gap in an intrinsic semiconductor as the hole and electron densities are equal. When the semiconductor is doped the carrier densities change, consequently the position of the Fermi level also changes. The shift in the position of the Fermi level can easily be related to the majority carrier density in a doped semiconductor if it is assumed that the addition of impurities do not affect the densities of energy states in the conduction and valence bands.

If N_c is the density of states in the conduction band and N_v the density of states in the valence band, we have for an intrinsic semiconductor

$$n = N_c \exp \frac{E_{fi}-E_c}{kT}$$

$$p = N_v \exp \frac{E_v-E_{fi}}{kT}$$

where E_{fi} is the energy associated with the Fermi level in an intrinsic semiconductor.

As $n = p$ for an intrinsic semiconductor we have from the above expressions

$$\frac{N_c}{N_v} = \exp \frac{E_v+E_c-2E_{fi}}{kT}$$

If E_{fn} is the energy associated with the Fermi level in an N-type semiconductor having an electron density n we have

$$n = N_c \exp \frac{E_{fn}-E_c}{kT}$$

$$p = N_v \exp \frac{E_v-E_{fn}}{kT}$$

and

$$\frac{n}{p} = \frac{n^2}{n_i^2} = \frac{N_c}{N_v} \cdot \exp \frac{2E_{fn}-E_c-E_v}{kT}$$

$$= \exp \frac{2(E_{fn}-E_{fi})}{kT}$$

$$n = n_i \exp \frac{E_{fn}-E_{fi}}{kT}$$

For a P-type semiconductor (1.12)

$$p = n_i \exp \frac{E_{fi}-E_{fp}}{kT}$$

The shift in the Fermi level is generally associated with a potential. As E_{fn} and E_{fi} are electron energies in eV we can associate a potential ϕ so that it represents the shift in the Fermi level from the mid band or intrinsic level. The potential $|\phi| = \dfrac{E_{fn}-E_{fi}}{q}$

On substitution in expression (1.12) we have

$$n = n_i \exp \frac{q\phi_n}{kT}$$

and
$$|\phi_n| = \frac{kT}{q} \,.\, \ln \frac{n}{n_i}$$

Similarly for a P-type semiconductor

$$|\phi_p| = \frac{kT}{q} \ln \frac{p}{n_i} \tag{1.13}$$

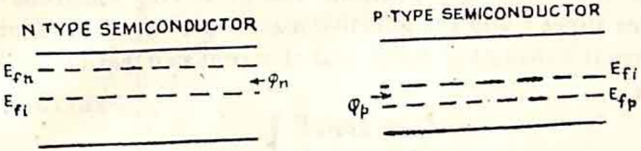


Fig. 1.4.

1.3.1 *Current flow in semiconductors*

The charge carriers, holes and electrons are in random motion within the semiconductor due to thermal vibrations. When a potential difference is applied to the ends of the semiconductor the charge carriers are acted on by a force due to the electric field and a bulk motion is superimposed on the random motion of the charge carriers. This bulk motion of the charge carriers due to the electric field is called drift.

The relation between the electric field and the velocity of bulk motion of the charge carriers involves the study of the random motion. Instead of considering motion of individual charge carriers we can consider their bulk motion. The charge carriers may be assumed to travel in straight lines between collisions and lose all their momentum at a collision. If $\bar{t}$ is the mean free time, a charge carrier suffers $\delta t/\bar{t}$ collisions in a time δt. The number of collisions suffered by N charge carriers in a time δt is therefore $N\,(\delta t/\bar{t})$. The charge carriers gain momentum from the field and lose it in collisions. Assuming that the charge carriers have a momentum p_x along the X axis, the momentum lost in time δt by the charge carriers in collisions is $Np_x\,\delta t/\bar{t}$. The momentum change $\delta p_x = (Np_x\,\delta t/\bar{t})$. Equating the rate of change in momentum to the force on the charge carriers due to the electric field we have

$$-NqE = \frac{\delta p_x}{\bar{t}} = \frac{-Np_x}{\bar{t}}$$

$$p_x = qE\bar{t} \tag{1.14}$$

If v_x is the average velocity per charge carrier the average momentum $p_x = m'v_x$, where m' is the effective mass of the charge carrier.

From the above expression we obtain the relation

$$v_x = qE\bar{t}/m' \tag{1.15}$$

The factor $q\bar{t}/m'$ is called the mobility of the charge carrier (μ) and we have the relation

$$v_x = \mu.E \tag{1.16}$$

The mobilities are different for holes and electrons as the mean free times $\bar{t}$ and the effective masses are not the same. The drift current densities of holes and electrons can easily be written down as

$$\left.\begin{aligned} J_p &= qp\mu_p E \\ J_n &= qn\mu_n E \end{aligned}\right\} \tag{1.17}$$

where p and n are densities of holes and electrons and μ_p, μ_n their mobilities. The conductivity of a sample of semiconducting material is obtained as

$$\sigma = q(\mu_p p + \mu_n n) \tag{1.18}$$

If the semiconductor is doped, the current carried by the minority carriers is negligible and we have the conductivities of *N*- and *P*-type samples

$$\sigma_n = q\mu_n n$$
$$\sigma_p = q\mu_p p$$

1.3.2 *Variation of mobility with field strength hot electron effect*

As stated above the electrons gain momentum from the field and lose it to the crystal lattice during collisions, energy is transferred from the field to the lattice. Under thermal equilibrium conditions the energy transferred during collisions is very small and the temperature of the crystal is not changed. The temperature corresponding to the random velocity of the electron called the electron temperature is the same as the crystal temperature. However at high fields the electron is accelerated to high velocities and the electron temperature is higher than the crystal temperature, the electron is referred to as a hot electron. Since the electron is not in thermal equilibrium with the lattice the electron velocity is no longer proportional to the field intensity. It is found that at high fields the electron velocity reaches a constant value of 10^5 met/sec called the scattering limited velocity. The relation between velocity and the field strength is shown in Fig. 1.5.

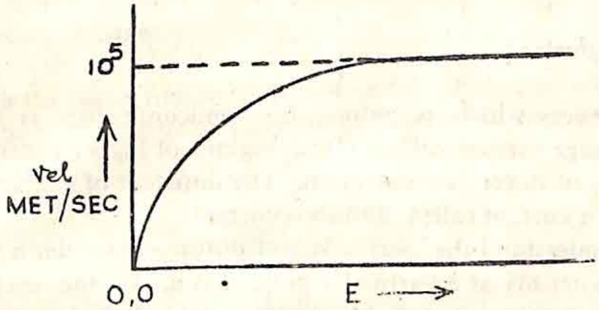


Fig. 1.5.

1.3.3 *Temperature dependence of mobility*

The mobility of a charge carrier is found to be directly proportional to the mean free time $\bar{t}$. The temperature dependence of mobility therefore is related to the scattering mechanisms in the crystal and their effect on the mean free time. Two types of scattering are present in a semiconductor—lattice scattering and impurity scattering. The lattice scattering represents the interaction of the electron with the lattice vibrations also referred to as an electron-phonon interaction. The mean free time for electron-phonon interactions depends on the thermal velocity of the electron and as the velocity increases with increase in temperature the mean free time decreases. The mobility arising out of lattice scattering decreases with increase in temperature. It is found that the mobility due to lattice scattering μ_l varies as $T^{-3/2}$.

The impurities present in the semiconductor are ionised and the ionised impurities deflect the charge carriers. The mobility due to impurity scattering μ_i depends on the time a charge carrier spends near an ionised impurity, as such μ_i is inversely proportional to the velocity of the charge carrier. μ_i is found to be proportional to $T^{3/2}$. The resultant mobility due to lattice and impurity scattering mechanisms can be written down as

$$\frac{1}{\mu}=\frac{1}{\mu_l}+\frac{1}{\mu_i} \tag{1.19}$$

The mobility μ is controlled by lattice scattering at high temperature and by impurity scattering at low temperatures. Even though it is theoretically proved that the mobility varies at $T^{-3/2}$, experimentally it is found to vary as $T^{-1.66}$ and $T^{-2.33}$ for N and P-type germanium and as T^{-1} and $T^{-2.1}$ for N and P-type silicon.

1.3.4 *Diffusion*

A process which is unique for semiconductors is diffusion where charge carriers diffuse from regions of higher concentration to regions of lower concentration. The diffusion of charge carriers results in a current called diffusion current.

To understand the mechanism of diffusion consider a situation where the density at a particular point A is n. As the carriers are randomly moving it can be assumed that half the number of carriers are moving to the right and half the number to the left. The same happens at all points and the number of particles moving through unit area is zero if the concentration is uniform. If the concentration is different at different points there is a net flow. To find the number of carriers moving through unit area in unit time consider two sections A' and A'' on either side of A at a distance equal to the mean free path 'l'. The distance is chosen as the mean free path so that we can assume that the particles moving from A' and A'' move through A without suffering a collision. If n_1 and n_2 are the densities at A' and A'' the number of particles crossing unit area of the section A in unit time is $\frac{1}{2}n_2v_{th}$ towards the left and $\frac{1}{2}n_1v_{th}$ towards right, where v_{th} is the thermal velocity. The net flow per second or the diffusion rate$=\frac{1}{2}(n_1-n_2)\cdot v_{th}$ writing

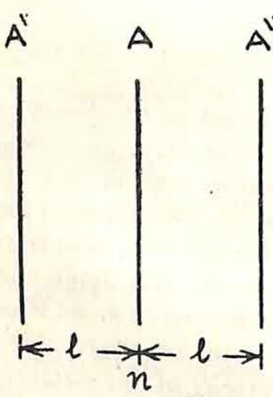


Fig. 1.6.

$$n_1=n-\frac{\partial n}{\partial x}\cdot l$$

and $$n_2=n+\frac{\partial n}{\partial x}\cdot l$$

The diffusion rate $=-l\cdot v_{th}\dfrac{\partial n}{\partial x}$ (1.20)

The rate of diffusion is proportional to the concentration gradient and the constant of proportionality lv_{th} is called the diffusion constant D. The minus sign indicates that the carriers diffuse from regions of higher density to regions of lower density. The diffusion constant

$$D=lv_{th}$$

If $\bar{t}$ is the mean free time $v_{th}=l/\bar{t}$ and $D=l^2/\bar{t}$. The units of D are met²/sec. The diffusion constant can be related to the mobility. We have found earlier that $\mu=\dfrac{q\bar{t}}{m'}$

$$\frac{D}{\mu}=\frac{lv_{th}}{q\bar{t}}=\frac{1}{q}\cdot m'\cdot v_{th}^2$$
$$=kT/q \tag{1.21}$$

as $m'v_{th}^2=kT$.

The above equation is called the Einstein's equation. The diffusion current densities for electrons and holes can now be written down as

$$J_n=qD_n\cdot\nabla n$$
$$J_p=-qD_p\cdot\nabla p$$

D_n and D_p are the diffusion constants for holes and electrons.

1.4.1 *Carrier generation and recombination*

Two physical processes which are of great importance in the operation of semiconductor devices are carrier generation and carrier recombination. It was stated earlier that an electron in the valence band moves into the conduction band leaving behind a hole in the valence band when it acquires an energy equal to the band gap. Electrons in crystals continuously move from the valence band to the conduction band deriving the necessary energy from thermal vibrations. This type of carrier generation is called thermal generation. Electrons also move into the conduction band from valence band when they receive energy in the form of light. The generation of carriers by light is called optical generation which will be discussed in detail under opto-electronic devices. Alongside carrier generation there is an inverse process called carrier recombination. Under thermal equilibrium conditions the generation rate g is equal to the recombination rate r and the carrier densities remain constant. The recombination rate or the number of carriers recombining per second in unit volume is proportional to the product of the densities of electrons and holes. The recombination rate

$$r = \alpha np \tag{1.22}$$

when α is a constant.

If n_0 and p_0 are the densities of electrons and holes under thermal equilibrium conditions the recombination rate equals the generation rate, and

$$r = \alpha n_0 p_0 = g \tag{1.23}$$

1.4.2 *Radiative recombination*

Radiative or direct recombination is one in which an electron recombines directly with a hole in the valence band. During the recombination an amount of energy equal to the difference between the electron and hole energies is emitted in the form of light. As momentum is to be conserved during recombination direct or radiative recombination occurs when the electron and hole momenta are numerically the same. This happens in a direct band gap semiconductors. If the hole and electron densities in a semiconductor are raised by Δn and Δp at time $t = 0$ the recombination rate in the semiconductor exceeds the thermal generation rate and the carrier densities return to equilibrium conditions exponentially with time. The density of the charge carriers at any subsequent instant of time can be found as follows.

The net recombination rate for electrons in the sample

$$\begin{aligned} dn/dt &= \alpha_n\,(n_0+\Delta n)\,(p_0+\Delta p)-g \\ &= \alpha_n(n_0+\Delta n)(p_0+\Delta p)-\alpha_n(p_0 n_0) \\ &= \alpha_n(n_0\Delta p+p_0\Delta n+\Delta n\Delta p) \end{aligned}$$

For low injection, i.e., when the increase in electron and hole densities are small compared to the thermal equilibrium values $\Delta n\cdot\Delta p$ can be neglected and

$$dn/dt = \alpha_n(n_0\Delta p+p_0\Delta n) \tag{1.24}$$

Similarly the net recombination rate for holes in the sample is

$$dp/dt = \alpha_p(n_0\Delta p+p_0\Delta n) \tag{1.25}$$

If the sample under consideration is an extrinsic semiconductor, the fractional change is small in the majority carrier density and it can be assumed that the majority carrier density remains constant. The change in minority carrier density with time is important and we have

$$dn/dt = \alpha_n p_0 . \Delta n \text{ (for a } P\text{-type semiconductor)}$$

and $$dp/dt = \alpha_p n_0 . \Delta p \text{ (for an } N\text{-type semiconductor)}$$

From the above equations we can write the expressions for minority carrier densities as functions of time

$$\delta n = \Delta n \exp \frac{-t}{\tau_n} \tag{1.26}$$

$$\delta p = \Delta p \exp \frac{-t}{\tau_p} \tag{1.27}$$

where $\tau_n = \alpha_n p_0$ and $\tau_p = \alpha_p n_0$; τ_n and τ_p are called radiative life times for electrons and holes respectively.

1.4.3 *Non-radiative recombination*

The electrons from the conduction band even when they move to the valence band may not emit energy in the form of light. The energy released instead of being emitted as light is absorbed by another electron and lost to phonons. This type of recombination is called the Auger process.

The most common recombination process in a semiconductor is by way of energy states in the forbidden gap introduced by impurities and crystal imperfections. An energy state within the forbidden gap may receive an electron from the conduction band and the electron falling into the energy state may return to the conduction band after some time. When this happens the energy state is called an electron trap as the electron is temporarily removed from the conduction band. Similarly a filled energy state within the forbidden gap may lose an electron temporarily to the valence band, when this happens a hole is removed from the valence band. The filled energy state is a hole trap. If the electron from the electron trap instead of returning to the conduction band moves into the valence band, a hole is removed and the energy state is referred to as a recombination centre.

The kinetics of recombination was worked out by Shockley and Read on the assumption of existence of single energy states with energy $E(t)$ within the forbidden gap. An electron in the conduction band moves into the energy state when it approaches the state, as such we can associate an area σ_n, called capture cross section for the electron. If v_t is the mean thermal velocity of electrons, an electron sweeps out a volume $\sigma_n v_t$ in one second. It falls into a trap if a trap is located in the volume. The number of electrons falling into traps in one second or the capture rate R_c is the product of the number of vacant traps in a volume $\sigma_n v_t$ and the density of electrons in the conduction band. The density of vacant traps is given by the expression $N(t)[(1-F(t))]$, where $N(t)$ is the density of traps and $F(t)$ the Fermi probability function for the trap level

$$F(t) = \frac{1}{1+\exp\dfrac{E_t-E_f}{kT}} \tag{1.28}$$

The capture rate

$$R_c = N(t)[1-F(t)]\sigma_n v_t n \tag{1.29}$$

The electrons in the traps are also in random motion and they return to the conduction band when a vacant energy state is available within a volume $\sigma_n v_t$. the number of electrons returning to the conduction band in one second or the rate of emission R_e is the product of the density of electrons in the trap levels $N(t)F(t)$ and the number of vacant energy states in a volume $\sigma_n v_t$ of the conduction band. If $N(c)$ is the density of states within the conduction band $F(c)$ is the Fermi probability function for the conduction band. The density of vacant energy states in the conduction band is $N(c)\,[1-F(c)]$ and the rate of emission

$$R_e = N(t)\,F(t)\,N(c)\,[1-F(c)]\,\sigma_n v_t \tag{1.30}$$

Under thermal equilibrium conditions the rates of emission and capture are the same and $n = n_0$. From expressions (1.29) and (1.30) we find

$$\begin{aligned} N(c)\,[1-F(c)] &= n_0 \cdot \frac{1-F(t)}{F(t)} \\ &= n_0 \cdot \exp\frac{E_t-E_f}{kT} \end{aligned} \tag{1.31}$$

Using expression (1.12) we have

$$N(c)\,[1-F(c)] = n_i \exp\frac{E_t-E_i}{kT}$$

When the electron density in the conduction band is more than the equilibrium value there is net capture of electrons by traps and the net capture rate for electrons is

$$R_n = N(t)\sigma_n v_t\left[(1-F(t))\,n - F(t)\,n_i \exp\frac{E_t-E_i}{kT}\right] \tag{1.32}$$

Similarly the difference between the holes captured by traps and the holes emitted from traps in one second or the net capture rate for holes can be found as

$$R_p = N(t)\sigma_p v_t \left\{ pF(t) - [1 - F(t)]\, n_i \exp \frac{E_i - E_t}{kT} \right\} \tag{1.33}$$

As the holes and electrons are recombining via traps the net capture rates for electrons and holes are the same. Equating R_p and R_n we have

$$F(t) = \frac{\sigma_n n + \sigma_p \cdot n_i \exp \dfrac{E_i - E_t}{kT}}{\sigma_p\left(p + n_i \exp \dfrac{E_i - E_t}{kT} \right) + \sigma_n \left(n + n_i \exp \dfrac{E_t - E_i}{kT} \right)} \tag{1.34}$$

Substituting the value of $F(t)$ from expression (1.34) in (1.32), we find that the recombination rate

$$R = \frac{N(t)\ \sigma_p \sigma_n v_t\ (pn - n_i^2)}{\sigma_p\left[p + n_i \exp \dfrac{E_i - E_t}{kT} \right] + \sigma_n \left[n + n_i \exp \dfrac{E_t - E_i}{kT} \right]} \tag{1.35}$$

The recombination rate is maximum when the trap level is situated at the middle of the band gap. The maximum recombination rate is therefore obtained by setting $E_t = E_i$

$$R = \frac{N(t)\ \sigma_p \sigma_n\ (pn - n_i^2)}{\sigma_p\ (p + n_i) + \sigma_n (n + n_i)} \tag{1.36}$$

For an N-type semiconductor if Δn is the excess carrier density we have

$$n = n_0 + \Delta n$$
$$p = p_0 + \Delta p$$

and $\Delta p = \Delta n$ as the charge neutrality is maintained. Substituting the values of n and p and neglecting the terms in p_0 in comparison to terms in n_0 and n_i as

$p_0 < n_i \ll n_0$ we have

$$R = N(t)\ \sigma_p v_t \cdot \frac{n_0 \Delta p}{(n_0 + \Delta n)}$$

For small injection, i.e. when the excess carrier density is small in comparison with the majority carrier density

$$R = N(t)\ \sigma_p v_t \cdot \Delta p$$

The minority and majority carrier densities decrease due to recombination. As the excess carrier density is small compared to the majority carrier density the fractional change in majority carrier density is negligible compared to the majority carrier density. The rate of recombination of minority carriers is proportional to the excess minority carrier density. The ratio of the minority carrier density to the rate of recombination

$$\frac{\Delta p}{R}=\frac{1}{N(t)\,\sigma_p v_t} \tag{1.37}$$

where $\frac{1}{N(t)\,\sigma_p v_t}$ has the dimensions of time and is referred to as the life time of holes in an *N*-type semiconductor τ_p. The minority carrier life time in an *N*-type semiconductor

$$\tau_p=\frac{\Delta p}{R}=\frac{p_n-p_{no}}{dp/dt}$$

or

$$\frac{dp}{dt}=\frac{p_n-p_{no}}{\tau_p} \tag{1.38}$$

where p_n is the density of holes or p_{no} is the thermal equilibrium minority density.

Similarly for a *P*-type semiconductor, the life time of electrons

$$\tau_n=\frac{n_p-n_{po}}{dn/dt}$$

and

$$\frac{dn}{dt}=\frac{n_p-n_{no}}{\tau_n} \tag{1.39}$$

Life time plays an important part in the operation of semiconductor devices. In some devices particularly high gain transistors it is necessary to use materials having long life times while in devices used for switching purposes the life times should be short. As the life time is inversely proportional to the density of recombination centres it can be adjusted to the required value by controlling the density of recombination centres. Long life times are achieved by keeping the impurities and crystal imperfections to a minimum. It is difficult to have long life times in low resistivity material as the density of impurities in low resistivity material is high. When short life times are required recombination centres are created by impurities purposely introduced. Gold introduced in silicon provides recombination centres at the middle of the band gap and the life times are considerably reduced.

1.5 Surface Recombination

Carriers recombine not only in the bulk of the semiconductor but also at the surface. The recombination at the surface of the semiconductor is due to recombination centres appearing at the surface. The density of recombination centres at the surface depends on surface treatment. When surface recombination occurs carriers diffuse from the bulk towards the surface and recombine. The surface recombination current is proportional to the difference between the density of the carriers at the surface and the equilibrium density. If J_p is the recombination current density due to holes recombining at the surface

$$\frac{1}{q} J_p \propto (p_s - p_0)$$

where p_s is the density of holes at the surface and p_0 the equilibrium value in the bulk.

Or

$$\frac{1}{q} J_p = s\,(p_s - p_0) \tag{1.40}$$

where s is a constant. The constant s has the dimensions of velocity and is referred to as the surface recombination velocity. Similarly for electrons we can write

$$\frac{1}{q} J_n = -s\,(n_s - n_0) \tag{1.41}$$

The surface recombination also follows the Shockley-Read mechanism as in the case of bulk recombination.

1.5.1 *Continuity equation*

The carrier densities within a semiconductor change due to recombination and current flow. The time rate of change of carrier density at a point in the semiconductor can easily be found in terms of the current and the life time. The expression for the time rate of change in carrier density is called the continuity equation. If we consider unit volume of a semiconductor having a hole density p, we find that some of the holes are lost due to recombination and the holes moving out of the volume constitute the hole current. The time rate of change in hole density at a point is the sum of the recombination rate of holes and the rate of hole flow due to the current. If p is the hole density at a point and p_0 the equilibrium density in the semiconductor the rate of recombination is $-\frac{p-p_0}{\tau_p}$. The number of holes moving out of unit volume in unit time or the rate of hole flow out of the volume is $\frac{1}{q}\nabla\cdot J_p$, where

S.C.E.R.T, W.B. LIBRARY
Date 23.12.98
Accn. No. 9787

J_p is the current density. The time rate change in hole density is now obtained as

$$\frac{dp}{dt}=\frac{-(p-p_0)}{\tau_p}-\frac{1}{q}\nabla\cdot J_p \tag{1.42}$$

J_p, the current density, is due to drift and diffusion, as such it can be written as

$$J_p = q\mu_p pE - qD_p\nabla p \tag{1.43}$$

Considering a one-dimensional case and substituting the value of J_p from expression (1.43) into (1.42) we obtain

$$\frac{dp}{dt}=-\frac{p-p_0}{\tau_p}-\mu_p p\cdot\frac{dE}{dx}-\mu_p E\frac{dp}{dx}+D_p\cdot\frac{d^2p}{dx^2} \tag{1.44}$$

If it is assumed that the space charge neutrality is maintained within the semiconductor the field terms disappear and we have

$$\frac{dp}{dt}=-\frac{p-p_0}{\tau_p}+D_p\cdot\frac{d^2p}{dx^2} \tag{1.45}$$

Similarly for electrons we find that

$$\frac{dn}{dt}=-\frac{n-n_0}{\tau_n}+D_n\cdot\frac{d^2n}{dx^2} \tag{1.46}$$

1.5.2 *Carrier density profile—diffusion length*

When the excess density is maintained constant at a point in a semiconductor the carriers move away from the point due to diffusion and there is also recombination. The continuity equation has to be solved for evaluating the excess density at different points in the semiconductor. If the excess density at a point $x = 0$ is maintained at a constant value δp_1 we have $dp/dt = 0$ and the continuity equation for holes becomes

$$0=\frac{-\delta p}{\tau_p}+D_p\frac{d^2p}{dx^2}$$

Solving the equation we have

$$p = C\exp\frac{-x}{\sqrt{D_p\tau_p}}$$

Writing $\sqrt{D_p\tau_p} = L_p$

$$p = C\exp\frac{-x}{L_p}$$

As $\delta p = \delta p_1$ at $x = 0$, $C=\delta p_1$

and $$\delta p = \delta p_1 \exp \frac{-x}{L_p} \tag{1.47}$$

Similarly for electron

$$\delta n = \delta n_1 \exp \frac{-x}{L_n}, \text{ where } L_n = \sqrt{D_n \tau_n}$$

The excess density decreases exponentially with distance according to the expression (1.47). L_p is the distance at which the excess density falls to $1/e$ times the value at $x = 0$ and it is referred to as the diffusion length. The diffusion length depends on the life time and the diffusion constant, when the life time of the charge carriers is very small the diffusion length is small.

Important Formulae in the Chapter

Effective mass of an electron

$$\frac{\hbar^2}{\partial^2 E/\partial k^2}$$

Fermi function

$$\frac{1}{1+\exp \frac{E-E_f}{kT}}$$

Carrier densities. Doped semiconductors

$$n = n_i \cdot \exp \frac{q\phi_n}{kT}$$

$$p = n_i \cdot \exp \frac{-q\phi_p}{kT}$$

$$\phi_n = \frac{kT}{q} \ln \frac{n}{n_i}$$

$$\phi_p = \frac{kT}{q} \ln \frac{p}{n_i}$$

ϕ_n and ϕ_p are the shifts in Fermi level.

Conductivity of a semiconductor material

$$q\,(n\mu_n + p\mu_p)$$

Einstein's relation

$$\frac{D}{\mu} = \frac{kT}{q}$$

Recombination rate

$$R = \frac{N(t)\,\sigma_p\sigma_n v_t\,(pn - n_i^2)}{\sigma_p\left(p + n_i \exp \frac{E_i - E_t}{kT}\right) + \sigma_n\left(n + n_i \exp \frac{E_t - E_i}{kT}\right)}$$

$$\approx N_t\sigma_p \cdot v_t \cdot \Delta p \text{ for holes}$$
$$\approx N_t\sigma_n \cdot v_t \cdot \Delta n \text{ for electrons}$$

Life times

$$\tau_p = \frac{p_n - p_{no}}{dp/dt} \quad \text{for holes}$$

$$\tau_n = \frac{n_p - n_{po}}{dn/dt} \quad \text{for electrons}$$

Surface recombination

$$\frac{1}{q} J_p = s(p_s - p_0)$$
$$\frac{1}{q} J_n = -s(n_s - n_0)$$

s is the recombination velocity.

Continuity equations

$$\frac{dp}{dt} = -\frac{p - p_0}{\tau_p} + D_p \cdot \frac{d^2p}{dx^2}$$
$$\frac{dn}{dt} = -\frac{n - n_0}{\tau_n} + D_n \cdot \frac{d^2n}{dx^2}$$

Carrier density profile

$$\delta p = \delta p_1 \exp(-x/L_p)$$
$$\delta n = \delta n_1 \exp(-x/L_n)$$
$$L_p = \sqrt{D_p\tau_p} \cdot L_n = \sqrt{D_n\tau_n}$$

Problems

1. Distinguish between a direct band gap semiconductor and an indirect band gap semiconductor.
2. What is meant by effective mass? Evaluate it for an electron in a band and explain why electrons in narrow bands do not take part in electric conduction.
3. What factors lead to the conclusion that a vacant energy state has properties of a particle?
4. Derive an expression relating the shift in the Fermi level and the carrier density in an extrinsic semiconductor.
5. On what factors does the mobility of an electron depend? What is a hot electron?
6. Derive Einstein's relation between mobility and diffusion constant.

7. Distinguish between radiative and non-radiative recombination in a semiconductor?
8. What is meant by surface recombination velocity? On what factors does the surface recombination depend?
9. Prove that the rate of recombination is proportional to the excess carrier density (Shockley Read Theory).

2. Junctions

A junction of *P* and *N* semiconductors has some very interesting and useful properties and semiconductor junctions are the basis of several devices. In this chapter we discuss in detail the properties of junctions. A brief mention of different fabrication methods is also given.

2.1 Fabrication of Junctions

P-N junctions are formed in three different ways: alloying, diffusion and ion implantation.

In alloying, a third group impurity, indium is placed on an *N*-type semiconductor wafer and is heated in an atmosphere of hydrogen to about 160°C. At this temperature indium melts and the impurity atoms penetrate the *N*-type semiconductor wafer overcompensating the semiconductor. The semiconductor is changed to *P*-type to the extent the impurity atoms penetrate. On cooling we have a crystal having sharply divided *P* and *N* regions. The junction is abrupt. This is one of the early methods of fabrication.

The diffusion process is one of the modern methods of fabrication as it is the one used for manufacturing integrated circuits. The process enables fabrication of a large number of devices simultaneously. In this process *N*-type silicon is heated to about 1000°C in an atmosphere containing a high concentration of third group impurity normally boron. At this high temperature boron atoms diffuse into the semiconductor overcompensating it. The junctions formed by this process are graded as the change from *P*-type to *N*-type occurs gradually within a short distance.

In ion implantation third group impurity ions are accelerated to high velocities and are made to strike an *N*-type semiconductor slab. The impurity atoms penetrate the semiconductor changing it to *P*-type. The implantation can be done up to different depths by controlling the velocity of the striking ions. It is also possible to implant impurities at different depths. This technology is being developed and interesting results are expected.

2.2 Band Diagram of a *P-N* Crystal—Barrier Voltage

The band diagram indicates the top edge of the valence band and the lower edge of the conduction band, the edges of the bands are horizontal in the absence of an electric field indicating the fact that the energy of the electron does not change from point to point within the semiconductor. The band diagrams for *P* and *N* semiconductors are given in the last chapter. To draw the band diagram of a *P-N* crystal we make use of the fact that there cannot be any current in an isolated junction. As the current in a semiconductor can be due to drift and diffusion the current density equations are

$$\left.\begin{aligned} J_n &= q\mu_n nE + qD_n \text{ grad } n \\ J_p &= q\mu_p pE - qD_p \text{ grad } p \end{aligned}\right\} \tag{2.1}$$

The current densities can be zero only when the two components of the current are equal and opposite. (The diffusion component has a definite value as the concentration of charge carriers changes across the junction.) For a one-dimensional case we can write

$$\left.\begin{aligned} 0 &= \mu_n nE + D_n \cdot (dn/dx) \\ 0 &= \mu_p pE - D_p \cdot (dp/dx) \end{aligned}\right\} \tag{2.2}$$

The densities of electrons and holes at different points in the crystal depend on the energy difference between the Fermi level E_f and the mid band energy E_i. From the expression (1.12) we have for electrons

$$n = n_i \exp \frac{E_f - E_i}{kT} \tag{1.12}$$

and

$$\frac{dn}{dx} = \frac{n}{kT}\left(\frac{\partial E_f}{\partial x} - \frac{\partial E_i}{\partial x}\right) \tag{2.3}$$

Substituting the value of dn/dx from expression (2.3) in expression (2.2) we have after making use of the Einstein's relation, $D = \mu kT/q$

$$E + \frac{1}{q}\left(\frac{\partial E_f}{\partial x} - \frac{\partial E_i}{\partial x}\right) = 0 \tag{2.4}$$

The potential ϕ at any point in the crystal is related to the energy by the equation, $\phi = (-E_i/q)$ as such $1/q \cdot (\partial E_i/\partial x) = -(\partial\phi/\partial x)$ and is equal to the field intensity E. Writing E for $1/q \cdot (\partial E_i/\partial x)$ we find that for an isolated P-N crystal

$$\frac{1}{q} \cdot \frac{\partial E_f}{\partial x} = 0 \tag{2.5}$$

We can therefore conclude that the Fermi level is invariant in a P-N crystal under the thermal equilibrium conditions.

Having concluded that the Fermi level is invariant we can draw the band diagram for the P-N crystal thus.

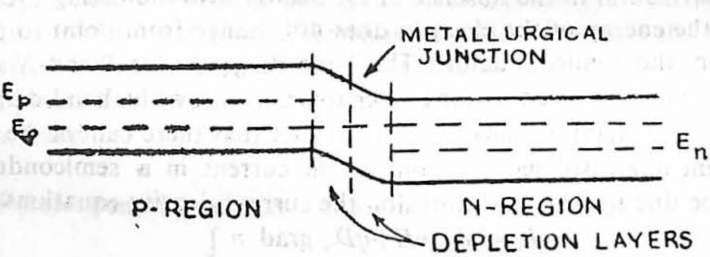


Fig. 2.1.

As the P-N junction is isolated there is no current and hence there is no bulk movement of the charge carriers in the P and N regions. Therefore there is no electric field in the bulk semiconductor regions on either side of the junction. This enables us to draw the bands horizontal in the bulk regions. Within narrow regions on either side of the metallurgical junction the bands are bent indicating the existence of an electric field. Due to the electric field the mobile carriers, holes and electrons are moved away exposing the fixed impurity ions. The region is referred to as the depletion layer or depletion region. Further calculations are made on the approximation that there are no mobile carriers in the depletion region.

The bands on the N side are depressed relative to the bands on the P side. This represents existence of a potential difference between the N and P regions. This potential difference is referred to as barrier voltage, diffusion voltage or contact potential. This barrier voltage is not measurable by a voltmeter because when a voltmeter is connected to measure the barrier voltage, contact potentials, at the points where contacts are made, neutralise the barrier voltage. The barrier voltage can however be calculated from the energy difference between the mid band energies on the N and P sides. The energy difference between the Fermi level and the mid band depends on the doping level. Equation (1.12) gives the relation between the energy difference and the doping level.

We obtain from expression (1.12) the energy difference between the Fermi level and the mid band energy E_n on the N side as

$$E_n - E_f = kT \ln \frac{n_n}{n_i} = kT \ln \frac{N_d}{n_i} \tag{2.6}$$

where N_d is the donor density on the N side of the crystal.

Similarly

$$E_f - E_p = kT \ln \frac{p_p}{n_i} = kT \ln \frac{N_a}{n_i} \tag{2.7}$$

where N_a is the acceptor density on the P-side. The energy difference between the mid-band energies on the P and N sides can be obtained from expressions (2.6) and (2.7) as

$$E_p - E_n = -kT \ln \frac{N_d N_a}{n_i^2} \tag{2.8}$$

The barrier voltage

$$V_b = \frac{-(E_p - E_n)}{q} = +\frac{kT}{q} \ln \frac{N_d N_a}{n_i^2} \tag{2.9}$$

For a one-sided abrupt junction when one side is heavily doped

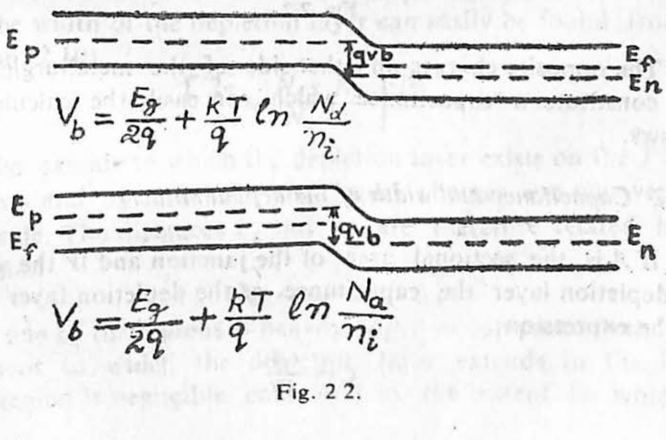

E_p
qV_b
E_f
E_n
$V_b = \frac{E_g}{2q} + \frac{kT}{q} \ln \frac{N_d}{n_i}$
E_p
E_f
qV_b
E_n
$V_b = \frac{E_g}{2q} + \frac{kT}{q} \ln \frac{N_a}{n_i}$

Fig. 2 2.

it can be assumed that the Fermi level coincides with the band edge on the heavily doped side. The band diagrams of one-sided step junctions and the barrier voltages are given in Fig. 2.2.

The barrier voltage calculated above exists across the depletion layer and it depends on the doping levels.

2.3.1 *Depletion layer*

The depletion layer consists of ionised impurity atoms on either side of the metallurgical junction. Even though the mobile carriers flow through the depletion layer the concentration of the mobile carriers at any point within the depletion layer is negligibly small compared to the impurity concentration and we can assume the region to be completely depleted. The impurity concentration changes abruptly at the metallurgical junction for a step junction while it changes rapidly from acceptor to donor in a graded junction. In either case we have space charges on either side of the metallurgical junction as indicated in Fig. 2.3.

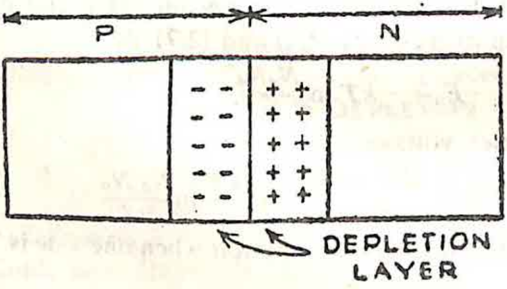


Fig. 2 3.

The opposite charges on either side of the metallurgical junction constitute a capacitance which can easily be calculated as follows.

2.3.2 *Capacitance and width of the depletion layer*

If A is the sectional area of the junction and W the width of the depletion layer the capacitance of the depletion layer is given by the expression

$$C=\frac{\epsilon A}{W} \tag{2.10}$$

where ϵ is the permittivity of the material.

When the voltage across the depletion layer changes there is a change in the width of the depletion layer and additional charges appear due to the widening of the space charge region. If δV is the change in the voltage the additional charge on the capacitor is $C\delta V$ and as it is due to widening of the space charge regions we can write for an abrupt junction

$$Q=C\delta V=AqN_d \cdot \delta x_n=AqN_a \cdot \delta x_p$$

where N_d and N_a are the densities of donors on the N side and acceptors on the P side; δx_n and δx_p are distances through which the depletion layer spreads on the N and P sides. The total increase in the width of the depletion layer

$$\delta W=\delta x_n+\delta x_p=\frac{C\delta V}{Aq}\left(\frac{1}{N_d}+\frac{1}{N_a}\right)$$

$$\frac{dW}{dV}=\frac{C}{Aq}\left(\frac{1}{N_d}+\frac{1}{N_a}\right) \tag{2.11}$$

Differentiating expression (2.10) with respect to V we have

$$\frac{dC}{dV}=-\frac{A\epsilon}{W^2}\cdot\frac{dW}{dV}$$

Substituting the value of $\frac{dW}{dV}$ from expression (2.11)

$$\frac{dC}{dV}=\frac{-C\epsilon}{W^2 q}\left(\frac{1}{N_d}+\frac{1}{N_a}\right) \tag{2.12}$$

Substituting the value of W from expressions (2.10) and (2.12) and integrating we have

$$C=A\left[\frac{q\epsilon}{2}\left(\frac{N_d\,N_a}{N_d+N_a}\right)\right]^{1/2} V^{-1/2} \tag{2.13}$$

The width of the depletion layer can easily be found from expression (2.10)

$$W=\left[\frac{2\epsilon}{q}\cdot\frac{N_d+N_a}{N_d\,N_a}\right]^{1/2}\cdot V^{1/2} \tag{2.14}$$

The extents to which the depletion layer exists on the P and N sides x_p and x_n are such that equal charges are uncovered on either side. The distances x_p and x_n are therefore related by the expression

$$N_d \cdot x_n=N_a \cdot x_p$$

If one of the regions is heavily doped as compared to the other, the extent to which the depletion layer extends in the heavily doped region is negligible compared to the extent to which the depletion layer extends in the less doped region. Junctions in which one side is heavily doped are referred to one-sided junctions and are written down as P^+N or PN^+ junctions, plus sign indicating the heavily doped region. For one-sided junctions we can assume that the depletion layer exists on one-side of the metallurgical junction.

For a PN^+ junction $N_d \gg N_a$ and expressions (2.13) and (2.14) reduce to

$$C = A\left(\frac{q\epsilon N_a}{2V}\right)^{1/2} \tag{2.15}$$

$$W = \left(\frac{2\epsilon V}{qN_a}\right)^{1/2} \tag{2.16}$$

The maximum field intensity within the depletion layer can easily be calculated in terms of the doping densities by using the Poisson's equation. Applying the Poisson's equation to the depletion layer on the N side we have

$$\frac{\partial^2 V}{\partial x^2} = \frac{-\rho}{\epsilon} = \frac{-qN_d}{\epsilon} \tag{2.17}$$

as the charge density in the space charge region on the N-side is qN_d. Integrating the expression we have

$$E = -\frac{dV}{dx} = \frac{qN_d x}{\epsilon} + \text{constant} \tag{2.18}$$

The value of the constant in the above equation is found using the fact that there is no electric field in the bulk semiconductor. If the width of the depletion layer on the N-side is x_n the constant in equation (2.18) is $(-qN_d \cdot x_n)/\epsilon$.

The field intensity increases linearly from zero value at the edge

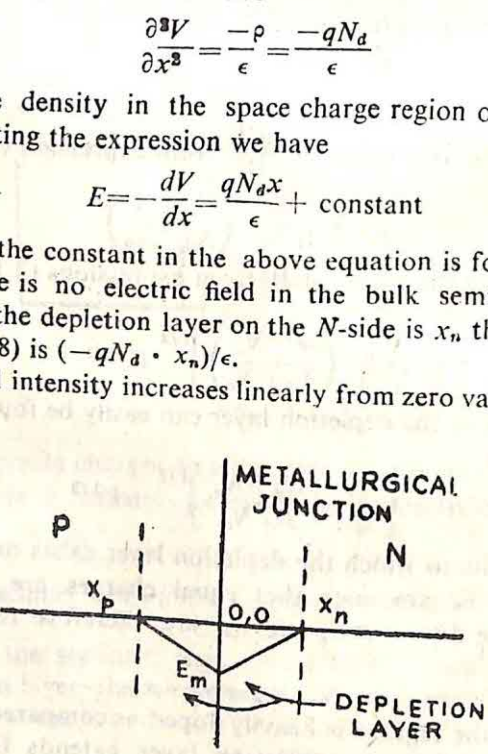


Fig. 2.4.

of the depletion layer to a maximum value at the metallurgical junction. The maximum value of the electric field

$$E_m=\frac{-qN_d\cdot x_n}{\epsilon} \tag{2.19}$$

The minus sign indicating the fact that the field is directed from N-side to the P-side (negative X direction).

As the field is increasing linearly with distance we can easily find the potential drop across the depletion region on the N-side as the product of the average field and the width of the depletion layer on the N-side

$$V_n=\frac{E_m}{2}\cdot x_n \tag{2.20}$$

Similarly the potential drop across the depletion layer on the P-side

$$V_p=\frac{E_m}{2}\cdot x_p \tag{2.21}$$

The voltage across the junction is the sum of the potential drops on the P and N sides as such

$$V=V_n+V_p=\frac{E_m}{2}(x_n+x_p)$$

$$=\frac{E_m}{2}\cdot W$$

The maximum field intensity

$$E_m=\frac{2V}{W}$$

Substituting the value of W from expression (2.14) we find that

$$E_m=\left[\frac{8\epsilon}{q}\,\frac{N_d+N_a}{N_dN_a}\right)^{1/2}\cdot V^{3/2} \tag{2.22}$$

For a one-sided junction

$$E_m=\left(\frac{8\epsilon}{qN_a}\right)^{1/2}\cdot V^{3/2} \tag{2.23}$$

2.3.3 *Graded junctions*

The junctions fabricated by modern techniques are all graded as diffusion processes are used in the fabrication When impurities are diffused at high temperature the impurity concentration in the crystal follows an error function. However, for ease in analysis, we assume the concentration to vary linearly. Due to

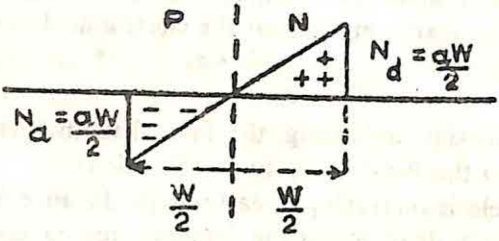


Fig. 2.5.

linear variation of impurity concentration the depletion layer is symmetrical about the metallurgical junction.

Let us consider a case where the gradient of the concentration is 'a' (units of a are met^{-4}) so that the charge density at a distance x from the metallurgical junction is qax. Applying the Poisson's equation we have

$$\frac{d^2V}{dx^2} = \frac{-qax}{\epsilon} \tag{2.24}$$

Integrating the above expression and making use of the fact that the field becomes zero at the edges of the depletion layer we find that the field intensity at a distance x from the metallurgical junction is given by the expression

$$E = \frac{qax^2}{2\epsilon} - \frac{qaW^2}{8\epsilon} \tag{2.25}$$

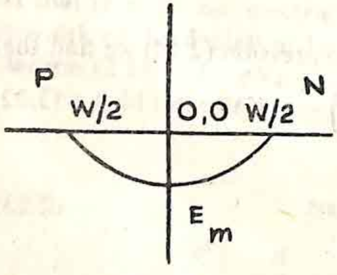


Fig. 2.6.

The field intensity within the depletion layer is plotted as a function of distance in Fig 2.6.

The field intensity reaches a maximum value at the metallurgical junction and the maximum value is given by the expression

$$E_m = \frac{-qaW^2}{8\epsilon} \tag{2.26}$$

The barrier voltage can be found in terms of a and W by integrating expression (2.25) between the limits $-W/2$ and $+W/2$ as

$$V_b = \frac{qaW^3}{12\epsilon} \tag{2.27}$$

The donor and acceptor densities at the outer edges of the depletion layer are $N_d = N_a = aW/2$ and the barrier voltage can also be obtained in terms of the impurity densities using the expression as

$$V_b = -\frac{kT}{q}\ln\frac{N_d\,N_a}{n_i^2}$$

$$= -\frac{2kT}{q}\ln\frac{aW}{2n_i} \tag{2.28}$$

Eliminating W from expressions (2.27) and (2.28) we obtain

$$V_b = -\frac{2kT}{q}\ln\frac{a}{2n_i}\left\{\frac{12\epsilon V_b}{qa}\right\}^{1/3} \tag{2.29}$$

The solution of the above equation gives the value of the barrier voltage. The equation is solved and a graph is drawn showing the relation between barrier voltage and carrier density gradient.

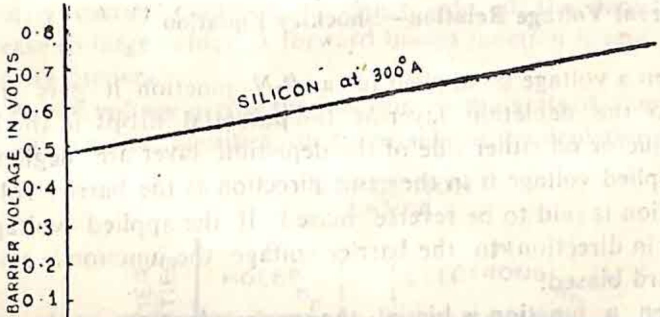


Fig. 2.7.

2.3.4 *Capacitance of a graded junction*

The capacitance of the graded junction can be easily obtained by writing $N_d = N_a = aW/2$ in expression (2.12). After the substitution we have

$$\frac{dC}{dV} = \frac{-4\epsilon C}{W^3 qa} \tag{2.30}$$

Substituting the value of W from the expression $C = A\epsilon/W$ we have

$$\frac{dC}{dV}=\frac{-4C^4}{qaA^3\epsilon^2}$$

Integrating the above expression we have

$$C=A\left(\frac{qa\epsilon^2}{12}\right)^{1/3}\cdot V^{-1/3} \tag{2.30}$$

Alternatively we can find the capacitance using the formula

$$C=\frac{A\epsilon}{W}$$

and substituting the value of W from expression (2.27)

$$C=A\left(\frac{qa\epsilon^2}{12}\right)^{1/3}\cdot V^{-1/3} \tag{2.30}$$

It is illustrative to note that the capacitance of an abrupt junction varies inversely as the square root of the voltage while the capacitance of a graded junction varies inversely as the cube root of the voltage.

2.4 Current Voltage Relation—Shockley Equation

When a voltage is applied to a *P-N* junction it gets transferred to the depletion layer as the potential drops in the bulk semiconductor on either side of the depletion layer are negligible. If the applied voltage is in the same direction as the barrier voltage the junction is said to be reverse biased. If the applied voltage is opposite in direction to the barrier voltage the junction is said to be forward biased.

When a junction is biased the carrier densities at the outer edges of the depletion layer are changed and the variation in the carrier densities can be found from the expression (2.9) which is applicable to an unbiased junction. If p_{po}, n_{po} are the densities of holes and electrons on the P side and p_{no}, n_{no} the densities of holes and electrons on the N side we have $N_d=n_n$, $N_a=p_p$ and $n_i^2=p_{po}n_{po}=p_{no}n_{no}$.

On substitution in equation (2.9) we find that

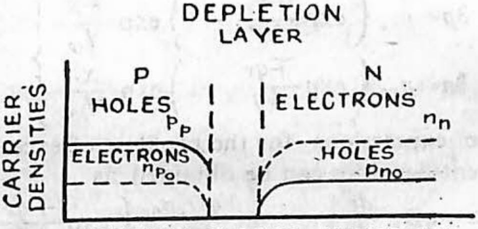
$$\left.\begin{aligned} n_{po}&=n_{no}\exp\frac{-qV_b}{kT}\\ \text{and}\qquad p_{no}&=p_{po}\exp\frac{-qV_b}{kT}\end{aligned}\right\} \tag{2.31}$$

When the junction is biased the voltage across the depletion layer becomes V_b+v for reverse bias and V_b-v for forward bias. The change in junction voltage causes a change in the majority and minority carrier densities on either side of the junction. For small applied voltages the change in the majority carrier density is negligible compared to the majority carrier density and it can be assumed that the majority carrier density remains constant. The minority carrier densities on either side of the depletion layer are given by the expressions

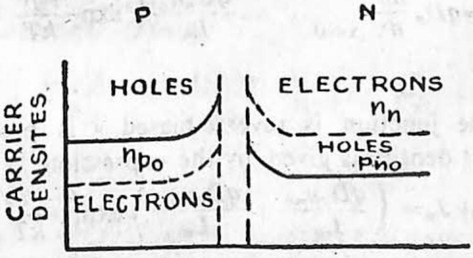
$$\left.\begin{aligned} n_p &= n_{no} \exp \frac{-q(V_b \pm v)}{kT} = n_{po} \exp \frac{\mp qv}{kT} \\ p_n &= p_{no} \exp \frac{\mp qv}{kT} \end{aligned}\right\} \qquad (2.32)$$

When the junction is reverse biased the exponential term becomes almost zero even for a small reverse voltage as kT/q is about 26 mV at room temperature, consequently the minority carrier densities on either side of the depletion layer fall to almost zero values. A reverse biased junction is said to collect minority carriers. When the junction is forward biased v is negative and the minority carrier densities on either side of the depletion layer increase to large values. A forward biased junction is said to inject minority carriers.

As the voltage across the junction is maintained constant the minority carrier densities on either side of the depletion layer are



REVERSE BIASED JUNCTION



FORWARD BIASED JUNCTION

Fig. 2.8.

maintained at values other than equilibrium values. The minority carrier densities return to their equilibrium values at a distance from the edge of the depletion layer due to diffusion and carrier generation or recombination. The carrier density profile in the presence of diffusion and recombination was derived in the last chapter. The excess hole density at a distance x from the edge of the depletion layer on the n side and the electron density at a distance $-x$ from the edge of the depletion layer or the p side are given by the expressions

$$\left.\begin{aligned} \delta p &= \delta p_1 \exp \frac{-x}{L_p} \\ \delta n &= \delta n_1 \exp \frac{x}{L_n} \end{aligned}\right\} \qquad (1.47)$$

The excess hole and electron densities at the edges of the depletion layer, δp_1, and δn_1 can be found from expression (2.32) as

$$\left.\begin{aligned} \delta p_1 &= p_n - p_{no} = p_{no}\left(\exp \frac{\mp qv}{kT} - 1\right) \\ \delta n_1 &= n_p - n_{po} = n_{po}\left(\exp \frac{\mp qv}{kT} - 1\right) \end{aligned}\right\} \qquad (2.33)$$

Substituting the values of δp_1 and δn_1 from expression (2.33) into (1.47) we obtain

$$\left.\begin{aligned} \delta p &= p_{no}\left(\exp \frac{\mp qv}{kT} - 1\right) \exp \frac{-x}{L_p} \\ \delta n &= n_{po}\left(\exp \frac{\mp qv}{kT} - 1\right) \exp \frac{x}{L_n} \end{aligned}\right\} \qquad (2.34)$$

From the expressions for the profiles of excess densities the diffusion current densities can be obtained as

$$J_p = -qD_p \left.\frac{dp}{dx}\right|_{x=0} = \frac{qD_p p_{no}}{L_p}\left(\exp \frac{\mp qv}{kT} - 1\right) \qquad (2.35)$$

$$J_n = qD_n \left.\frac{dn}{dx}\right|_{x=0} = -\frac{qD_n n_{po}}{L_n}\left(\exp \frac{\mp qv}{kT} - 1\right) \qquad (2.36)$$

Reverse bias

When the junction is reverse biased v is positive and the reverse current density is given by the expression

$$J_r = J_n + J_p = \left(\frac{qD_p p_{no}}{L_p} + \frac{qD_n n_{po}}{L_n}\right) \exp\left(\frac{-qv}{kT} - 1\right)$$

for large values of the reverse voltage $\exp \frac{-qv}{kT}$ can be neglected in comparison with 1 and we have

$$J_r=\left(\frac{qD_p\,p_{no}}{L_p}+\frac{qD_n\,n_{po}}{L_n}\right) \tag{2.37}$$

The reverse current according to the above expression is independent of the voltage and it is generally referred to as reverse saturation current. The reverse current shows a small increase with voltage due to carrier generation within the depletion layer and surface effects which we will be discussing in a later section.

Forward bias

When the junction is forward biased the minority carrier densities on either side of the depletion layer increase to large values and δp_1 and δn_1 have values

$$\left.\begin{aligned}\delta p_1&=p_{no}\left(\exp\frac{qv}{kT}-1\right)\\ \delta n_1&=n_{po}\left(\exp\frac{qv}{kT}-1\right)\end{aligned}\right\} \tag{2.38}$$

The carrier density profile for a forward bias is given by the expressions

$$\left.\begin{aligned}\delta p&=p_{no}\left(\exp\frac{qv}{kT}-1\right)\exp\frac{-x}{L_p}\\ \delta n&=n_{po}\left(\exp\frac{qv}{kT}-1\right)\exp\frac{x}{L_n}\end{aligned}\right\} \tag{2.39}$$

The diffusion current densities can now be obtained from expression (2.39) as

$$\left.\begin{aligned}J_p&=-qD_p\cdot\frac{dp}{dx}\bigg|_{x=0}=\frac{qD_p\,p_{no}}{L_p}\left(\exp\frac{qv}{kT}-1\right)\\ J_n&=qD_n\cdot\frac{dn}{dx}\bigg|_{x=0}=\frac{qD_n\,n_{po}}{L_n}\left(\exp\frac{qv}{kT}-1\right)\end{aligned}\right\} \tag{2.40}$$

The forward current density

$$J=J_p+J_n=\left(\frac{qD_p\,p_{no}}{L_p}+\frac{qD_n n_{po}}{L_n}\right)\left(\exp\frac{qv}{kT}-1\right)$$

$$=J_r\exp\left(\frac{qv}{kT}-1\right) \tag{2.41}$$

The forward current therefore increases exponentially with the applied voltage. The expression (2.41) is not valid for small forward voltages due to recombination in the depletion layer.

2.5 Quasi Fermi Levels

We have noted above that the minority carrier densities within short distances on either side of the depletion layer are different from the thermal equilibrium values when a junction is biased. The condition that the product of the density of holes p and the density of electrons n in a semiconductor is a constant and is equal to n_i^2 is not valid within short distances on either side of the depletion layer. The product is less than n_i^2 when the junction is reverse biased. It is more than n_i^2 when the junction is forward biased.

The densities of holes and electrons are related to the Fermi energy and the mid band energy by the expressions

$$\left.\begin{aligned} p&=n_i \exp\frac{E_i-E_f}{kT} \\ n&=n_i \exp\frac{E_f-E_i}{kT} \end{aligned}\right\} \tag{1.12}$$

The expressions can be applied to the semiconductor regions on either side of the depletion layer of the junction under bias by assuming that the Fermi level changes at the edges of the depletion layer to account for the change in minority carrier density. The Fermi levels are different for holes and electrons as the density of minority carriers are only affected by biasing. The Fermi levels governing the densities of carriers under nonthermal equilibrium conditions are referred to quasi-Fermi levels or Imrefs.

If δ is the energy difference between the quasi-Fermi level for holes and the Fermi level on the N-side, we have from expression (1.12)

$$\begin{aligned} p_n&=n_i \exp \frac{E_i-E_f+\delta}{kT} \\ &=p_{no} \cdot \exp \frac{\delta}{kT} \end{aligned} \tag{2.42}$$

Comparing expressions (2.33) and (2.42) we find that

$$\delta=\mp qv \tag{2.43}$$

The quasi-Fermi level for holes is shifted upwards by an amount equal to the applied potential when the junction is reverse biased. The quasi-Fermi level for electrons is shifted downwards on the P-side. The shifts are in opposite directions when the junction is forward biased.

We have found earlier that the hole and electron current den-

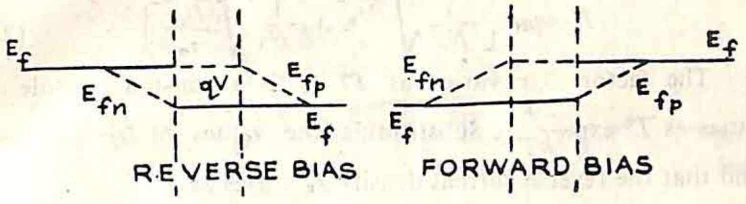


Fig. 2.9. Dashed lines represent quasi-Fermi levels and full lines represent Fermi levels.

sities in the isolated crystal are zero as the Fermi level is invariant. From expressions (2.1) we find that the current densities J_p and J_n when the junction is biased are given by the expressions

$$J_p=\mu_p \cdot p \cdot \frac{\partial E_{fp}}{\partial x}$$

$$J_n=\mu_n \cdot n \cdot \frac{\partial E_{fn}}{\partial x}$$

When the junction is biased the quasi-Fermi levels vary as indicated above. If a potential is associated with the quasi-Fermi level so that $q\phi=-E_f$ we find that the current densities under biased conditions are given by the expressions

$$J_p=-q\mu_p\, p \cdot \frac{\partial \phi_p}{\partial x}$$

$$J_n=-q\mu_n\, n \cdot \frac{\partial \phi_n}{\partial x}$$

where ϕ_p and ϕ_n are the potentials associated with the quasi-Fermi levels for holes and electrons.

2.6 Temperature Dependence of Reverse Saturation Current

The reverse saturation current density of a junction is given by the expression

$$J_r=q\left(\frac{D_p\, p_{no}}{L_p}+\frac{D_n \cdot n_{po}}{L_n}\right) \tag{2.37}$$

The minority carrier densities p_{no} and n_{po} are related to the doping densities by the expressions $p_{no}=\frac{n_i^2}{N_d}$, $n_{po}=\frac{n_i^2}{N_a}$.

Substituting the values of the minority carrier densities and the values of L_p and L_n we have

$$J_r = q n_i^2 \left[\frac{1}{N_d} \sqrt{\frac{D_p}{\tau_p}} + \frac{1}{N_a} \sqrt{\frac{D_n}{\tau_n}} \right] \tag{2.44}$$

The factor D/τ varies as T^{γ} (γ is a constant) while n_i^2 varies as $T^3 \exp\frac{-E_g}{kT}$. Substituting the values of D/τ and n_i^2 we find that the reverse current density J_r varies as

$$T^{\left(3+\frac{\gamma}{2}\right)} \exp \cdot \frac{-E_g}{kT}$$

The variation in the reverse current due to the term $T^{\left(3+\frac{\gamma}{2}\right)}$ is negligible compared to the variation due to the exponential term, we can say that the reverse current varies as $\exp\frac{-E_g}{kT}$

$$J_r \propto \exp\frac{-E_g}{kT} \tag{2.45}$$

The variation of J_r with temperature can be used to determine the band gap of the semiconductor.

2.7 Carrier Generation and Recombination within the Depletion Layer

We found in Chapter 1 that whenever there is a deviation from thermal equilibrium conditions there is either carrier recombination or carrier generation. The recombination rate is given by the expression

$$R = \frac{N(t)\sigma_p \, \sigma_n \, v_t \, (pn - n_i^2)}{\sigma_p\left(p + n_i \exp\frac{E_i - E_t}{kT}\right) + \sigma_n\left(n + n_i \exp\frac{E_t - E_i}{kT}\right)} \tag{1.35}$$

If we remember that the generation process is the reverse of the recombination process a similar expression gives the generation rate. In the depletion layer of a reverse biased junction the hole and electron densities are negligibly small as such p and n can be neglected. The generation rate is given by the expression

$$U = \frac{-n_i \sigma_p \sigma_n v_t}{\sigma_p \exp\frac{E_i - E_t}{kT} + \sigma_n \exp\frac{E_t - E_i}{kT}} = \frac{-n_i}{\tau_e} \tag{2.46}$$

where τ_e is the effective lifetime given by the expression

$$\tau_e = \frac{\sigma_p \exp \dfrac{E_i - E_t}{kT} + \sigma_n \exp \dfrac{E_t - E_i}{kT}}{\sigma_p \sigma_n v_{th}} \tag{2.47}$$

As the generation rate is the number of carriers generated in unit volume in unit time the current density due to carrier generation

$$J_g = q \mid U \mid W$$

where W is the width of the depletion layer.

Substituting the value of the generation rate

$$J_g = \frac{q\, n_i\, W}{\tau_e} \tag{2.48}$$

The generation current density in a reverse biased junction is proportional to n_i while the reverse diffusion current in an ideal diode is proportional to n_i^2. For germanium n_i^2 is very large, as such the generation current is negligibly small compared to the reverse diffusion current. However, for wide band gap semiconductors Si and GaAs n_i is small and the generation current cannot be neglected.

As the generation current is proportional to the width of the depletion layer W, it increases with the reverse voltage due to widening of the depletion layer. The reverse current therefore is not independent of the reverse voltage as expected in an ideal diode.

The ideal diode equation also fails under forward biased conditions as current is lost due to recombination within the depletion layer. The loss in the number of carriers due to recombination is to be found by integrating expression (1.35). The integration is complicated due to the fact that the hole and electron densities vary from point to point within the depletion layer. However it is possible to find the effect of recombination on the forward characteristic by assuming that the hole and electron densities remain constant within the depletion. In other words, we are assuming that the quasi-Fermi levels remain constant within the depletion layer. Expression (1.35) for the rate of recombination is considerably simplified by assuming that the trap levels are situated at the middle of the band gap and $\sigma_n = \sigma_p = \sigma$. With these assumptions the rate of recombination is given by the expression

$$R = \frac{N_t \sigma v_{th}\,(pn - n_i^2)}{p + n + 2n_i} \tag{2.49}$$

The hole and electron densities in the depletion layer can be expressed in terms of quasi-Fermi potentials as

$$n=n_i \exp\frac{q(\psi-\phi_n)}{kT}, \quad p=n_i \exp\frac{q(\phi_p-\psi)}{kT}$$

where ϕ_n and ϕ_p are the quasi-Fermi potentials and ψ the potential corresponding to the middle of the band gap. Substituting the values of p and n in expression (2.49) we have

$$R=\frac{N_t\cdot\sigma v_{th}\, n_i\left(\exp\frac{qV}{kT}-1\right)}{\exp\frac{q(\psi-\phi_n)}{kT}+\exp\frac{q(\phi_p-\psi)}{kT}+2} \tag{2.50}$$

The recombination rate is maximum when the sum $p+n$ is a minimum in the expression (2.49). As the product pn is a constant the sum is minimum when $p=n$, i.e. the recombination rate is maximum at the point where the intrinsic level is situated mid-way between the two quasi-Fermi levels. The maximum recombination rate can be found from expression (2.50) by writing $\psi=\frac{\phi_p+\phi_n}{2}$ when the expression reduces to

$$R_{\max}=\frac{N_t\cdot\sigma v_{th}\, n_i\left(\exp\frac{qV}{kT}-1\right)}{2\left(\exp\frac{qV}{2kT}-1\right)} \tag{2.51}$$

when $V \gg \frac{kT}{q}$ the above expression is further simplified

$$R_{\max}\approx\frac{N_t\sigma v_{th}\, n_i}{2}\cdot\exp\frac{qV}{2kT} \tag{2.52}$$

The recombination current density can be obtained as

$$J_{\text{rec}}=\int_0^W qR\,dx$$

$$\approx\frac{qN\sigma v_{th} n_i}{2}\cdot W\cdot\exp\frac{qV}{2kT} \tag{2.53}$$

The recombination current is also proportional to n_i just as the generation current in a reverse biased junction. For low forward voltages the current in silicon and other wide band semiconductor junctions varies as exp $(qV/2kT)$. The recombination for very low forward voltages is considerable and the current is not established till a minimum forward voltage is applied. The minimum forward voltage necessary to establish a current is called cut in voltage and it is quite important in the design of electronic circuits. The cut in voltage for silicon junctions is about 0.5 volts.

2.8 Junction Breakdown

The current through a reverse biased junction was found to be very small and fairly independent of the applied voltage. That is why it is referred to as reverse saturation current. When the reverse voltage is increased to a certain value it is found that the reverse current increases considerably and the rectifying action of the junction is lost, the junction is said to have broken down. Two different physical processes are involved in the breakdown. One of the processes is called Zener breakdown and another Avalanche breakdown. The two different breakdowns are used to advantage in certain circuit applications. The junction breakdown does not destroy the junction so long as the current is not excessive.

Zener breakdown

The slight increase in reverse current with voltage was found to be due to carrier generation within the depletion layer. The carriers are thermally generated within the depletion layer and as the generation rate is maintained constant the reverse current shows a small increase with voltage due to the widening of the depletion layer. As the depletion layer is very narrow the field intensity within the depletion layer increases considerably with the applied voltage. When the field intensity reaches a value of about 10^5 V/cm, the bound electrons in the depletion layer are field excited into the conduction band creating electron hole pairs. The electrons and holes generated within the depletion layer move in opposite directions under the influence of the intense electric field causing a sudden increase in the current. The current is limited only by the external circuit. This breakdown is referred to as Zener breakdown. The breakdown can also be viewed as a consequence of tunneling from valence band to conduction band. Since the width of the depletion layer is a function of the doping densities the field intensity may be made to reach the limiting breakdown value for different applied voltages and Zener diodes can be manufactured to breakdown at well-defined voltages. Zener diodes are invariably used in voltage regulating circuits.

Avalanche breakdown

For low doping densities the depletion layer is fairly wide and the field intensity is not sufficient to cause field emission. Under these conditions the junction may break down by impact ionisation. As this mechanism is important in transistors and in the study of certain microwave devices we will study the process in a little detail.

A charge carrier moving with a certain velocity within the semiconductor suffers a collision with the lattice. If the carrier has sufficient energy the collision results in the creation of an electron hole pair. We can define an ionisation rate α as the number of electron hole pairs generated by the carrier in travelling through unit distance. The ionisation rate is strongly dependent on the intensity of the electric field and is different for holes and electrons. The dependence of the ionisation rate on the field intensity is due to change in mean free path and allied considerations. We restrict ourselves to the effects of ionisation and not go into details of the variation of ionisation rate with field intensity

Consider an elementary slice AB of thickness δx within the depletion layer of a reverse biased junction. If I_p is the hole current entering the section from the left, the number of holes entering the section is I_p/q and the number of electrons hole pairs generated in the section is $\frac{I_p}{q} \cdot \alpha_p \cdot dx$, where α_p is the ionisation rate for holes. The increase in hole current due to ionisation within the section is $I_p \cdot \alpha_p \cdot dx$. Similarly if I_n is the electrons current entering the section from the right, the increase in the hole current caused by electrons is $I_n \cdot \alpha_n \cdot dx$, where α_n is the ionisation rate for electrons. The increase in hole current due to ionisation within the slice AB

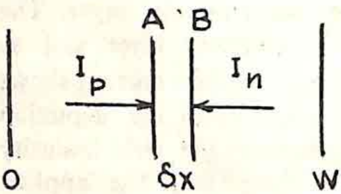


Fig. 2.10.

$$\delta I_p = \alpha_p I_p \cdot dx + \alpha_n I_n \cdot dx$$

As the total current $I = I_p + I_n$

$$\frac{dI_p}{dx} = \alpha_n I + (\alpha_p - \alpha_n) I_p \qquad (2.54)$$

To obtain the current multiplication due to impact ionisation expression (2.54) is to be integrated over the entire depletion layer. If W is the width of the depletion layer and I_{po} and $M \cdot I_{po}$ are the hole currents on either side of the depletion layer. We have after integrating the above expression and setting

$$I=M_p I_{po}$$

$$M_p I_{po}-I_{po}=M_p I_{po}\int_0^W \alpha_p \exp\left[-\int_0^x \left(\alpha_p-\alpha_n\right) dx'\right]dx$$

$$1-\frac{1}{M_p}=\int_0^W \alpha_p \exp\left[-\int_0^x \left(\alpha_p-\alpha_n\right) dx'\right]dx \tag{2.54}$$

The factor M_p is called the hole multiplication factor. The junction breaks down and the current reaches a very high value when the multiplication factor becomes infinite, i.e. when the integral in expression (2.54) becomes 1. The integral is referred to as the ionisation integral. The breakdown may be either due to ionisation initiated by holes or electrons. The ionisation integrals for holes and electrons are different and are given by expressions

$$\int_0^x \alpha_p \exp\left[-\int_0^x (\alpha_p-\alpha_n)dx'\right] dx \qquad \text{for holes} \tag{2.55}$$

$$\int_0^x \alpha_n \exp\left[-\int_0^x (\alpha_p-\alpha_n)dx'\right] dx \qquad \text{for electrons} \tag{2.56}$$

In certain semiconductors GaAs and GaP the ionisation rates are equal and the ionisation integral is simplified to the form

$$\int_0^W \alpha\, dx \tag{2.57}$$

The breakdown voltage of a junction can be found by evaluating the ionisation integral and setting it to one. The integral has to be evaluated using numerical methods and a computer as the ionisation rates are functions of electric field intensity which varies from point to point in the depletion layer

The breakdown voltage depends on the doping levels and on temperature. From a knowledge of the breakdown voltage the critical field at which breakdown occurs is evaluated. The current through a reverse biased junction shows an increase with the applied voltage for voltages near the breakdown voltage. The ratio of the current to the reverse saturation current, called the multiplication factor, M, is related to the voltage across the junction by an empirical relation

$$M = \frac{1}{1 - \left(\frac{V}{V_{br}}\right)^n} \tag{2.58}$$

where n has a value between 3 and 6 depending on the material.

2.9 Charge Storage—Admittance of a Junction

It has been proved earlier that the minority carrier densities on either side of the depletion layer increase to large values depending on the forward voltage. To preserve charge neutrality within the bulk semiconductor on either side of the depletion layer majority carriers flow into the semiconductor regions from the external circuit. The majority carriers flow out of the semiconductor regions when the minority carrier densities decrease. As such we find that there is a flow of majority carriers into and out of the semiconductor regions similar to that of a capacitance. A forward biased junction has a capacitance due to the excess density of minority carriers and the capacitance is called the storage capacitance.

To evaluate the storage capacitance of a junction we assume that a small alternating voltage of maximum value v is superimposed on a steady forward voltage V_0. The voltage across the diode at any particular instant is given by the expression

$$V = V_0 + v \exp j\omega t \tag{2.59}$$

The hole density on the N-side of a P^+N junction can now be written down as

$$p_n = p_{no} \exp \frac{q}{kT}\left(V_0 + v \exp j\omega t\right)$$

$$= p_{no} \exp \frac{qV_0}{kT} \exp \frac{qv}{kT}$$

(exp $j\omega t$ is dropped as the voltage v is understood to be alternating).

As the alternating voltage is small the exponential can be expanded to give

$$p_n = p_{no} \exp \frac{qV_0}{kT}\left(1 + \frac{qv}{kT}\right)$$

$$= p_{no} \exp \frac{qV_0}{kT} + p_{no} \exp \frac{qV_0}{kT} \cdot \frac{qv}{kT} \tag{2.60}$$

The excess hole density on the *N*-side is now obtained as

$$p_n-p_{no}=\delta p_n=p_{no}\left(\exp\frac{qV_0}{kT}-1\right)+p_{no}\exp\frac{qV_0}{kT}\cdot\frac{qv}{kT}$$

writing $$p_{no}\left(\exp\frac{qV_0}{kT}-1\right)=\delta\bar{p}_n$$

and $$p_{no}\exp\frac{qV_0}{kT}\cdot\frac{qv}{kT}=\delta\tilde{p}_n$$

we have $$\delta p_n=\delta\bar{p}_n+\delta\tilde{p}_n$$

The excess hole density on the *N*-side consists of a steady part $\delta\bar{p}_n$ and alternating part with a maximum value $\delta\tilde{p}_n$. The continuity equation for holes on the *N*-side can be split into two parts: an alternating part and a steady part. From the steady state equation we find that the excess hole density at a distance *x* from the edge of the depletion layer is given by the expression

$$\delta p=\overline{\delta p}_n\exp\frac{-x}{L_p}$$

and the steady state current density

$$J_p=-qD_p\cdot\left.\frac{dp}{dx}\right|_{x=0}$$

$$=\frac{qD_p}{L_p}\cdot p_{no}\left(\exp\frac{qV_0}{kT}-1\right)$$

As we are considering a P^+N junction the electron current density is negligible and the total current density *J* is the same as J_p

$$J=\frac{qD_p}{L_p}\cdot p_{no}\left(\exp\frac{qV_0}{kT}-1\right).$$

The steady state current

$$I=AJ=\frac{AqD_p}{L_p}\cdot p_{no}\left(\exp\frac{qV_0}{kT}-1\right)$$

The incremental conductance *g* of the junction is obtained from the above equation as

$$g=\frac{dI}{dV}=\frac{AqD_p}{L_p}\cdot p_{no}\frac{q}{kT}\exp\frac{qV_0}{kT} \qquad (2.61)$$

$$=\frac{q}{kT}\cdot I \qquad (2.62)$$

For the alternating part of the continuity equation we have $dp/dt=j\omega\tilde{p}_n$.

Substituting the value of $\frac{dp}{dt}$ we have

$$j\omega\tilde{p} = \frac{-\tilde{p}}{\tau_p} + D_p \cdot \frac{d^2\tilde{p}}{dx^2}$$

$$0 = -\tilde{p}\left(\frac{1}{\tau_p} + j\omega\right) + D_p \cdot \frac{d^2\tilde{p}}{dx^2}$$

writing $\frac{1}{\tau_p} + j\omega = \frac{1}{\tau'_p}$. The solution for the alternating part of the continuity equation can be found similar to the steady state case. The alternating excess hole density at a distance x from the edge of the depletion layer is

$$\tilde{\delta p} = \tilde{\delta p_n} \exp \frac{-x}{L'_p}$$

where $L'_p = \sqrt{D_p \tau'_p}$.

The alternating current density

$$\tilde{J}_p = -qD_p \left.\frac{d\tilde{p}}{dx}\right|_{x=0}$$

$$= \frac{qD_p}{L'_p} \cdot \tilde{\delta p_n}$$

$$= \frac{qD_p}{L'_p} \cdot p_{no} \cdot \exp \frac{qV_0}{kT} \cdot \frac{qv}{kT}$$

The alternating current flowing through the junction

$$i = \frac{AqD_p}{L'_p} \cdot p_{no} \cdot \exp \frac{qV_0}{kT} \cdot \frac{qv}{kT}$$

We now readily obtain the admittance as

$$Y = \frac{i}{v} = \frac{AqD_p}{L'_p} p_{no} \exp \frac{qV_0}{kT} \frac{q}{kT}$$

Substituting the value of $AqD_p p_{no} \cdot \exp \frac{qV_0}{kT} \cdot \frac{q}{kT}$ from equation (2.61) we have

$$Y = g \frac{L_p}{L'_p}$$

$$= g\sqrt{\frac{\tau_p}{\tau'_p}} = g\sqrt{1 + j\omega\tau_p} \qquad (2.63)$$

The root in expression (2.63) can be approximated as the alternating voltage is small compared to the steady voltage. We have, after approximation

$$Y = g + \frac{j\omega\tau_p}{2} \cdot g \tag{2.64}$$

The junction behaves as a parallel combination of a resistor of conductance g and *a* capacitance $\frac{g\tau_p}{2}$.

The capacitance is called the storage or diffusion capacitance

$$C_D = \frac{g\tau_p}{2} = \frac{q}{kT} \cdot I \cdot \frac{\tau_p}{2} \tag{2.65}$$

The capacitance of a forward biased junction is the sum of the storage and transition capacitances. As the transition capacitance is very small it is often neglected in comparison with the storage capacitance of the forward biased junction.

The stored charge contributing to the capacitance can be found by integration as

$$C = \frac{dQ}{dV} = \frac{q}{kT} \cdot I \cdot \frac{\tau_p}{2}$$

$$Q = \int_0^{V_0} \frac{q}{kT} \cdot I \cdot \frac{\tau_p}{2}$$

Expressing I in terms of V and integrating we find that

$$Q = \frac{I \cdot \tau_p}{2} \tag{2.66}$$

This is the charge that is associated with the capacitance, i.e. the charge that can be reclaimed. As we know that the charge is due to excess holes we can find the charge from the excess hole distribution on the *N*-side.

The hole density at a distance x from the depletion layer is given by the expression

$$\delta p_n = \delta p_{no} \exp - \frac{x}{L_p}$$

And the charge due to excess holes

$$Q = Aq\int_0^{\infty} p_n = Aq\delta p_{no} \int_0^{\infty} \exp \frac{-x}{L_p}$$

$$= \frac{Aq\delta p_{no}}{L_p}$$

$$= I\tau_p \tag{2.67}$$

as $$I = \frac{AqD_p}{L_p} \cdot \delta p_{no}$$

The reclaimable charge is only half the charge associated with the excess holes. The other half of the charge escapes through the junction. The above calculations are made on the assumption that the N-side is infinitely long so that the excess hole density falls to zero exponentially within the N-type semiconductor. However, if the N-side is not wide enough, i.e. for a narrow based diode it can be shown that two thirds of the change is reclaimable when the applied voltage falls to zero.

For simplification the calculations are made for a P^+N junction. The results for a PN junction are similarly obtained.

2.10 Transient Behaviour of a Diode

In this section we consider the effect of charge storage on the transient behaviour of a diode. When the voltage across a diode is suddenly reversed we expect the current to fall from a large value to the reverse saturation value instantaneously, the diode behaving like an ideal switch. The current does not change as in the case of an ideal switch due to charge storage. To understand the change in current with a sudden reversal of voltage consider a P^+N diode carrying a forward current I_f. The voltage across the diode is related to the excess hole density on the N-side by the expression (2.38). We can write

$$v \approx \frac{kT}{q} \ln \frac{\delta p_n}{p_{no}} \tag{2.68}$$

From the above expression we notice that a change in voltage also implies a change in excess density. The diode current cannot change from a forward value to the reverse saturation value as the excess hole density on the N-side cannot change immediately from a large value to zero value. During the interval the excess density is above the thermal equilibrium value and the diode has a small forward voltage across its terminals. the reverse voltage

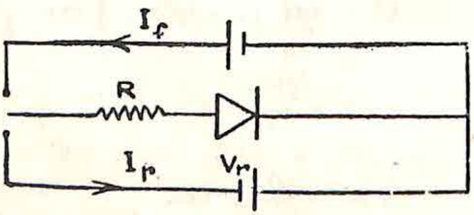


Fig. 2.11.

appearing across the resistance in series with the diode drives a current $I_r = V_r/R$.

The time during which a steady current I_r flows through the diode is called the storage time. The storage time can be roughly found as the ratio of the change reclaimable to the reverse current

$$t_s = \frac{Q_s}{2I_r}$$
$$= \tau_p \cdot \frac{I_f}{2I_r} \tag{2.69}$$

The storage time evaluated in the above manner is higher than the actual storage time as carrier recombination is not taken into account. The storage time was evaluated by Kingston taking into account recombination. The Kingston's expression is

$$t_{sd} = \tau_p \left(\text{erfc} \frac{I_f}{I_f + I_r} \right)^2 \tag{2.70}$$

when $I_r = I_f$ in magnitude, the error function approximately has a value 1/2 and the storage time works out as $\tau_p/4$. (The storage time under similar conditions is $\tau_p/2$ using the rough estimate.)

After the minority carrier density returns to the thermal equilibrium value the reverse voltage starts appearing across the diode and the current drops to the reverse saturation value. The time during which the current falls from I_r to 10% of its value is called the decay time and it is often much less than the storage time. The wave forms of the applied voltage and current through the diode are as shown in Fig. 2.12.

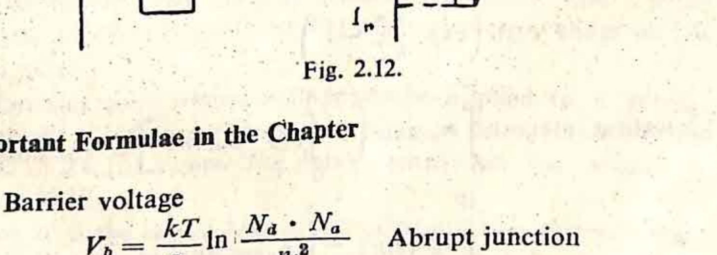


Fig. 2.12.

Important Formulae in the Chapter

Barrier voltage

$V_b = \frac{kT}{q} \ln \frac{N_d \cdot N_a}{n_i^2}$ Abrupt junction

$$V_b = \frac{2kT}{q} \ln \frac{a}{2n_i} \frac{(12\epsilon V_b)}{qa} \quad \text{Graded junction}$$

Width of the depletion layer

$$W = \left(\frac{2}{q} \cdot \frac{N_d + N_a}{N_d \, N_a} \right)^{\frac{1}{2}} V^{\frac{1}{2}} \quad \text{Abrupt junction}$$

$$= \left(\frac{2\epsilon V}{qN_a} \right)^{\frac{1}{2}} \quad PN^{+} \text{ junction}$$

$$= \left(\frac{2\epsilon V}{qN_d} \right)^{\frac{1}{2}} \quad P^{+}N \text{ junction}$$

$$= \left(\frac{12\epsilon V}{q_a} \right)^{1/3} \quad \text{Graded junction}$$

Junction capacitance

$$C = A \left[\frac{q\epsilon}{2} \cdot \frac{N_d \, N_a}{N_d + N_a} \right]^{\frac{1}{2}} \cdot V^{-\frac{1}{2}} \quad \text{Abrupt junction}$$

$$= A \left(\frac{q\epsilon N_a}{2V} \right)^{\frac{1}{2}} \quad PN^{+} \text{ junction}$$

$$= A \left(\frac{qA\epsilon^2}{12} \right)^{\frac{1}{2}} \cdot V^{-1/3} \quad \text{Graded junction}$$

Maximum field intensity in the depletion layer

$$E_m = \left[\frac{8\epsilon}{q} \cdot \frac{N_d + N_a}{N_d \, N_a} \right]^{\frac{1}{2}} \cdot V^{3/2} \quad \text{Abrupt junction}$$

$$= \left(\frac{8\epsilon}{qN_a} \right)^{\frac{1}{2}} \cdot V^{3/2} \quad PN^{+} \text{ junction}$$

$$= \frac{qaW^2}{8\epsilon} \quad \text{Graded junction}$$

Reverse saturation current density in a junction

$$J_r = q \left(\frac{D_p p_{no}}{L_p} + \frac{D_n n_{po}}{L_n} \right)$$

Current equation

$$I = I_0 \left(\exp \frac{qv}{kT} - 1 \right)$$

Ionisation integral $\int_0^W \alpha_p \exp \left[- \int_0^x (\alpha_p - \alpha_n) \, dx' \right] dx$ for holes

$$\int_0^W \alpha_n \exp \left[- \int_0^x (\alpha_p - \alpha_n) dx' \right] dx \quad \text{for electrons}$$

Avalanche breakdown condition $\int_0^W \alpha \cdot dx = 1$

Multiplication factor $M = \dfrac{1}{1-\left(\dfrac{V}{V_{br}}\right)^n}$

Incremental forward conductance $g = \dfrac{q}{kT} \cdot I$

Diffusion capacitance P^+N junction

$$C_D = \frac{q}{kT} \cdot I \cdot \frac{\tau_p}{2}$$

Storage time $t_{sd} = \tau_p \cdot \dfrac{I_f}{2I_r}$

Questions

1. Show that the Fermi level is invariant for an isolated junction. Draw the band diagram for a junction.
2. Derive an expression for the barrier voltage existing across a P-N junction. If the donor and acceptor densities on the N and P sides of a silicon junction are 10^{22} met^{-3}, evaluate the barrier voltage.
3. Draw the band diagram of a P^+N step junction. Evaluate the barrier voltage of a silicon P^+N junction if the donor density on the N side is 1.45×10^{22} met^{-3}.
4. Derive expressions for the capacitance and width of the depletion layer of an abrupt junction.
5. Evaluate the width of the depletion layer for a P^+N junction if the donor density on the N-side is 1.45×10^{20} met^{-3}, when (a) the junction is unbiased; (b) when a reverse voltage of 5 V is applied.
6. What maximum reverse voltage can be applied to a germanium P^+N junction if the junction breaks down for an electric field of 2×10^7 V/met. The donor density on the N-sides is 2.4×10^{21} met^{-3}.
7. Show that the capacitance of a graded junction varies inversely as the cube root of the voltage across the junction.

8. Evaluate the reverse saturation current density in a junction.
9. Derive the relation between current and voltage for a junction.
10. What do you understand by quasi-Fermi levels? Derive the expression for current density in terms of the quasi-Fermi levels.
11. Is the reverse current in a junction constant and independent of applied voltage? Briefly discuss the variation of reverse current with the applied voltage.
12. What is the effect of recombination on the forward current when it is very small?
13. Distinguish between Zener and Avalanche breakdowns.
14. Evaluate the admittance of a forward biased junction and hence calculate the storage capacitance of the junction.
15. What is the origin of stored base charge? Prove that only half the charge is reclaimable.
16. Discuss the switching behaviour of a junction bringing out clearly the various times involved in the switching process.

3. Transistor

Transistor is a three-terminal-semiconductor device capable of producing amplification like a triode. It is manufactured by melting a third group element on either side of an *N*-type semiconductor wafer or by melting a fifth group element on either side of a *P*-type semiconductor wafer.

In the first case the third group impurity overcompensates the *N*-type semiconductor on either side producing *P* regions on either side of *N*-type wafer while in the second case the fifth group impurity overcompensates the *P*-type semiconductor on either side producing *N* regions on either side. The crystal structures formed are *PNP* and *NPN*. The impurities introduced are fairly large in concentration and the areas over which the impurities are melted are different on either side. The necessity for this will be obvious as we proceed further. The smaller area side is known as emitter and the larger area side collector and the semiconductor wafer sandwiched between emitter and collector is known as base. In a *PNP* transistor the emitter and collector regions are of *P*-type and the base *N*-type while in an *NPN* transistor the emitter and collector regions are *N*-type and the base *P*-type.

The junction separating the emitter and base is called the emitter junction while the junction between the collector and base is called the collector junction.

3.1.1 *Current flow through a transistor*

In normal operation the emitter junction is forward biased while the collector junction is reverse biased. Considering a *PNP* structure the hole density in the base at the edge of the emitter base junction has a large value due to minority carrier injection while the hole density in the base at the edge of the collector base junction has a very small value. The difference in the densities of holes on either side of the base region sets up a concentration gradient within the base. Under the influence of the concentration gradient holes diffuse across the base region from the emitter junction to the collector junction. The hole current due to diffusion at the collector junction adds to the reverse saturation current of the reverse biased collector junction and the collector current changes.

We can separate out the various components of the current. The emitter current or the current flowing through the emitter junction is due to holes injected from emitter to base and electrons injected from base to emitter. The electron component of the emitter current flows between emitter and base. The collector current is the sum of the reverse saturation current of the collector junction and the current due to holes reaching the collector junction. The base current has three components—the electron component of the emitter current, the reverse saturation current of the collector junction and the current due to recombination within the base region. These components are individually very small. By applying Kirchoff's laws to the transistor we have

$$I_E = I_C + I_B \tag{3.1}$$

where I_E is the emitter current, I_C the collector current and I_B the base current. The voltage change needed to vary the emitter current is very small as the emitter base junction is forward biased and the power associated with the emitter junction is very small. The changes in emitter current cause changes in the collector current. As the resistance of a reverse biased junction is very large the changes in collector current are associated with considerable power.

3.1.2 *Current calculations*

The ratio of the hole current at the base emitter junction to the total emitter current is called the emitter efficiency γ, the ratio of the hole current at the base collector junction to the hole current at the base emitter junction the transport factor β and ratio of the collector current I_C to the hole current at the base collector boundary is called the collector multiplication factor δ. From the above definitions it follows that the collector current is related to the emitter current by the relation

$$I_C=\gamma\beta\delta I_E \tag{3.2}$$

The ratio of the collector current I_C to the emitter current is called the current gain α, and $\alpha=\gamma\beta\delta$.

It is instructive to calculate the values of γ, β and δ so that the physical behaviour of a transistor can be appreciated. The calculations on a practical transistor are involved as such an ideal transistor having a uniform cross section is considered. The junctions are assumed to be parallel and abrupt. With these assumptions the factors γ, β and δ can be calculated.

3.1.3 *Emitter efficiency*

The emitter efficiency was defined as the ratio of the hole current at the emitter junction to the emitter current. The hole current at the emitter junction is evaluated from the gradient of the hole density at the edge of the depletion layer on the emitter side of the base. If V_E is the forward voltage of the base emitter junction the hole density on the emitter side of the base is

$$p_1 = p_0 \exp \frac{qV_E}{kT} \tag{3.3}$$

The hole density at the edge of the depletion layer on the collector side of the base is given by the expression

$$p_2 = p_0 \exp \frac{-qV_C}{kT} \tag{3.4}$$

As V_C is very large $\exp \frac{-qV_C}{kT}$ is almost zero and the hole density at the edge of the depletion layer on the collector side of the base is taken as zero.

The hole density profile in the base region can be obtained by substituting the boundary conditions in the general expression for carrier density profile

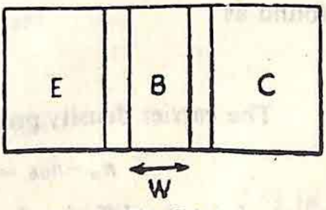


Fig. 3.1.

$$p_x = C_1 \exp \frac{x}{L_p} + C_2 \exp \frac{-x}{L_p} \tag{3.5}$$

At the edge of the depletion layer on the emitter side of the base $x = 0$, and $p_1 = p_0 \exp \frac{qV_E}{kT}$, at the edge of the depletion layer on the collector side of the base $x = W$ and $p_2 = 0$.

Substituting the boundary values in the general expression, the constants C_1 and C_2 are found as

$$C_1 = \frac{-p_1 \exp\left(-\frac{W}{L_p}\right)}{2 \sinh \frac{W}{L_p}} \tag{3.6}$$

$$C_2 = \frac{p_1 \exp\left(\frac{W}{L_p}\right)}{2 \sinh \frac{W}{L_p}} \tag{3.7}$$

The hole current density at the emitter base boundary is

$$\begin{aligned} J_p &= -qD_p \left.\frac{dp}{dx}\right|_{x=0} \\ &= \frac{-qD_p}{L_p}(C_1 - C_2) \\ &= \frac{qD_p}{L_p} p_1 \coth \frac{W}{L_p} \end{aligned} \tag{3.8}$$

The electron current through the base emitter junction can be found by considering the electron density profile on the *N*-side. The general expression for the carrier density profile has the following limits.

At $x = 0$ on the emitter side

$$n_1 = n_{0E} \exp \frac{qV_E}{kT} \tag{3.9}$$

Deep in the emitter side at $x = -\infty$, $n = n_{0E}$.

With these limits the constants in the profile expression can be found as

$$C_2 = 0$$
$$C_1 = n_1 - n_{0E}$$

The carrier density profile of electrons on the *P* side (emitter) is

$$n_x - n_{0E} = (n_1 - n_{0E}) \exp \frac{x}{L_n} \tag{3.10}$$

where L_n is the diffusion length of electrons in the emitter region.

The electron current density through the emitter junction

$$= qD_n \frac{dn}{dx}\bigg|_{x=0}$$

$$= qD_n \frac{n_1 - n_{0E}}{L_n} = \frac{qD_n\, n_{0E}\left(\exp \dfrac{qV_E}{kT} - 1\right)}{L_n} \tag{3.11}$$

The minus sign in the diffusion current equation cancels with the negative charge carried by the electron and the electron current at the junction is positive.

Neglecting 1 in comparison with $\exp \frac{qV_E}{kT}$ the electron current at the emitter base junction

$$= \frac{qD_n\, n_{0E}}{L_n} \exp \frac{qV_E}{kT}$$

The current density through the emitter junction

$$= \frac{qD_p}{L_p} p_0 \exp \frac{qV_E}{kT} \coth \frac{W}{L_p} + \frac{qD_n\, n_{0E}}{L_n} \exp \frac{qV_E}{kT} \tag{3 12}$$

The emitter efficiency is given by the expression

$$= \frac{\dfrac{qD_p}{L_p} p_0 \exp \dfrac{qV_E}{kT} \coth \dfrac{W}{L_p}}{\dfrac{qD_p}{L_p} p_0 \exp \dfrac{qV_E}{kT} \coth \dfrac{W}{L_p} + \dfrac{qD_n\, n_{0E}}{L_n} \exp \dfrac{qV_E}{kT}}$$

$$= \frac{1}{1 + \dfrac{qD_n\, n_{0E}}{L_n} \dfrac{L_p}{qD_p\, p_0} \tanh \dfrac{W}{L_p}} \tag{3.13}$$

The conductivity on the emitter side

$$\sigma_E = q\mu_p\, p_0$$

as the product of the hole density and electron density is a constant

$$\sigma_E = \frac{qD_p}{n_{0E}} \cdot \text{constant}$$

Similarly

$$\sigma_B = \frac{qD_n}{p_0} \cdot \text{constant}$$

$$\frac{\sigma_E}{\sigma_B} = \frac{D_p}{D_n} \cdot \frac{p_0}{n_{0E}} \tag{3.14}$$

Substituting the value of $\frac{\sigma_E}{\sigma_B}$, we have for the emitter efficiency

$$\gamma = \frac{1}{1+\frac{\sigma_B}{\sigma_E} \cdot \frac{L_p}{L_n} \tanh \frac{W}{L_p}} \tag{3.15}$$

In useful transistor structures the electrical base width W is much less than the diffusion length of holes in the base region L_p and

$$\tanh \frac{W}{L_p} \approx \frac{W}{L_p} \tag{3.16}$$

With this approximation the emitter efficiency

$$\gamma \approx \frac{1}{1+\frac{\sigma_B}{\sigma_B} \frac{W}{L_n}} \tag{3.17}$$

The emitter efficiency of a transistor as given by the above expression depends on the ratio of the conductivities of the base and emitter regions and also the ratio of the electrical base width and the diffusion length of electrons in the emitter region. The factor $\frac{\sigma_B}{\sigma_E} \cdot \frac{W}{L_n}$ must be very very small in order that the emitter efficiency is very close to 1. $\frac{\sigma_B}{\sigma_E}$ is made small by having a heavily doped emitter region. The factor $\frac{W}{L_n}$ can be reduced to a small value by making L_n large. The increase in L_n involves an increase in the lifetime of minority carriers in the emitter region, this is difficult because when the emitter region is heavily doped a large density of impurity states appear in the forbidden gap. However, the crystal imperfections and extraneous impurities can be reduced to a minimum so that the lifetime may be large enough to make W/L_n very small. In useful transistors the emitter efficiency is roughly 0.998 or 0.999.

3.1.4 *Transport factor*

The hole density profile in the base region is given by the expression

$$p_x = C_1 \exp \frac{x}{L_p} + C_2 \exp \left(\frac{-x}{L_p} \right)$$

The constants C_1 and C_2 are evaluated earlier and have values given by expressions (3.6) and (3.7).

The hole current density at the base collector junction is given by

$$-qD_p \frac{dp}{dx}\Big|_{x=W} = -\frac{qD_p}{L_p}\cdot\left[C_1\left(\exp\frac{W}{L_p}\right)-C_2\exp\left(-\frac{W}{L_p}\right)\right]$$

$$=\frac{qD_p}{L_p}\cdot\frac{p_1}{\sinh\frac{W}{L_p}} \tag{3.18}$$

The hole current density at the base emitter junction has been found as

$$\frac{qD_p}{L_p}\,p_1\coth\frac{W}{L_p} \tag{3.8}$$

The transport factor (β) is therefore

$$\beta=\operatorname{sech}\frac{W}{L_p} \tag{3.20}$$

$$\beta=1-\left(\frac{W}{L_p}\right)^2 \tag{3.21}$$

To make β very nearly equal to 1 it is necessary to make $\frac{W}{L_p}$ very very small. This is accomplished by decreasing W to a small value and increasing L_p in the base region by making the base region free from crystal imperfections.

3.1.5 *Collector multiplication factor*

The ratio of the collector current to the hole current at the base collector junction is known as collector multiplication factor (δ) as the electron current through the collector junction is negligible compared to the hole current, this factor has a value of 1 at normal operating conditions. However, when the collector voltage is increased there is an increase in collector current due to Avalanche multiplication at the collector junction. For large collector voltages the collector multiplication factor δ is given by the expression (2.58)

$$\delta = \frac{1}{1 - \left(\frac{V}{V_{br}}\right)^n} \tag{2.58}$$

where V is the applied reverse voltage across the collector junction and V_{br} the breakdown voltage, the factor n has a value between 3 and 6.

3.1.6 *Base spreading resistance*

The base current consisting of the electron currents flowing through the emitter and collector junctions of a *PNP* transistor and the current due to recombination of holes in the base region flows to the lead making contact with the base region of the transistor. The flow paths of the different components making up the base current are different. The components of the base current flow through the narrow base region to reach the contact as such they have to overcome certain resistance. Even though the components of the base current have different flow paths it is possible to consider a single resistance to be in series with the base, this resistance is called the base spreading resistance. It has a value of about 100 ohms.

3.2 D.C. Characteristics

Unlike a tube, a transistor is used in different ways. In a tube the voltages are always measured relative to the cathode and the current in the plate circuit is determined by the potential differences existing between the cathode and the grid, cathode and plate etc. A tube is generally operated with the grid at a negative potential and there is no grid current.

A transistor is a three-terminal device and can be operated with the base common to input and output, in this case it is said to be used in the common base configuration. A transistor can also be used with its emitter common, in which case the transistor is said to be used in the common emitter configuration. Another difference between a tube and transistor arises out of the fact that there are currents flowing in all the three leads. The currents are related to one another by the various formulae derived earlier. It is appropriate to say that while a tube is a voltage controlled device a transistor is a current controlled device. In the following pages we will discuss the various D.C. characteristics of a transistor first in the common base configuration and later in the common emitter configuration.

The sign convention for currents that is commonly used is to consider the currents flowing into the transistor as positive and the currents flowing out as negative. The voltages are always measured relative to the common lead. For a *PNP* transistor I_E is positive as I_E flows into the transistor. I_B and I_C are negative as they flow out of the transistor. The voltages V_{BE} and V_{CB} are negative.

3.2.1 *Common base configuration*

In the common base configuration the input voltage V_E is the forward voltage appearing across the emitter base junction and the output voltage V_C the reverse voltage across the collector base junction. The input current I_E is the forward current, I_C is the current flowing through the reverse biased collector junction.

Input characteristic

The input characteristic is the curve showing the relation between the input voltage and current for different values of collector base voltage. The current is given by the expression

$$I_E = Aq \left\{ \frac{D_p\, p_0 \exp \dfrac{qV_E}{kT}}{L_p} \coth \frac{W}{L_p} + \frac{D_n\, n_{0E}}{L_n} \exp \frac{qV_E}{kT} \right\}$$

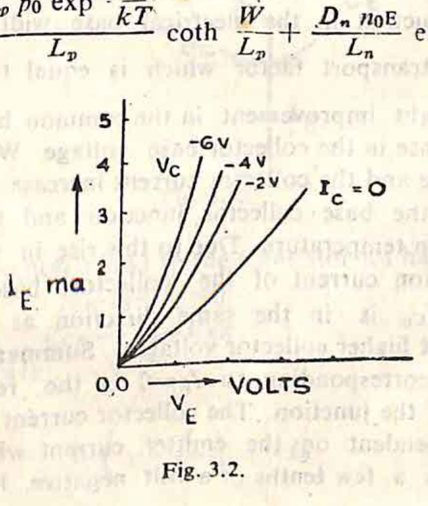


Fig. 3.2.

The variation of the emitter current I_E with emitter base voltage V_E is similar to a forward biased junction except that the gradient varies with the collector voltage, the change in the gradient with collector voltage arises from the narrowing of the base region due to the depletion layer at the collector base boundary moving into the base. When the collector is open the emitter current flows through the base only and the base spreading resistance which is much larger than the resistance of the forward biased diode comes in series, as a result the effective junction voltage is reduced and the input characteristic with collector open is much less steeper than the other characteristics.

Output characteristic

The output characteristics are the curves showing the relation between a collector voltage and collector current for different values of emitter current. In normal operation the collector base junction is reverse biased and the current voltage characteristic for zero emitter current is the same as that of a reverse biased diode. When the emitter current is increased the collector current reaches almost the same value as the emitter current when the collector is made a few tenths of a volt negative relative to the base. Under these conditions, as discussed earlier, the hole current leaving the emitter junction is collected by the collector. When the collector base voltage is increased the depletion layer of the collector base junction moves towards the emitter reducing the electrical base width. The reduction in the electrical base width results in an increase in the transport factor which is equal to $1-\left(\frac{W}{L_p}\right)^2$ There is a slight improvement in the common base current gain with the increase in the collector base voltage. When the reverse collector voltage and the collector current increase there is power dissipation at the base collector junction and there is a consequential rise in temperature. Due to this rise in temperature the reverse saturation current of the collector base junction I_{C_0} increases. As I_{C_0} is in the same direction as I_C there is an increase in I_C at higher collector voltages. Summarising the output characteristic corresponding to $I_E=0$ is the reverse saturation characteristic of the junction. The collector current reaches a constant value dependent on the emitter current when the collector base voltage is a few tenths of a volt negative. Large values of the collector voltage increase the collector current and the characteristics develop an upward bend.

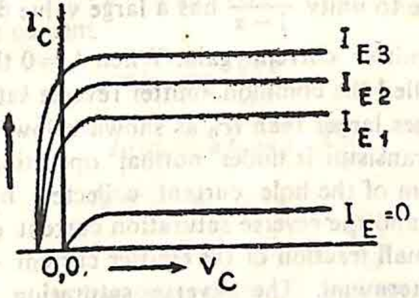


Fig. 3.3.

3.2.2 *Common emitter configuration*

In the common emitter configuration the emitter is common to the input and output.

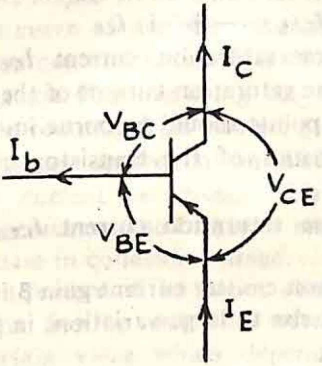


Fig. 3.4.

Applying Kirchoff's laws to the transistor we have

$$I_E+I_B+I_C=0 \tag{3.22}$$

$$V_{CE}=V_{CB}+V_{BE}$$

As $I_C=-\alpha I_E$

we have

$$I_C\left(1-\frac{1}{\alpha}\right)=-I_B$$

or
$$I_C = \frac{\alpha}{1-\alpha} \cdot I_B \tag{3.23}$$

As α is close to unity $\frac{\alpha}{1-\alpha}$ has a large value β, and β is called the common emitter current gain. When $I_B=0$ the collector current $-I_{CE0}$ is called the common emitter reverse saturation current. I_{CE0} is many times larger than I_{C0} as shown below.

When the transistor is under normal operation the collector current is the sum of the hole current collected by the collector junction $-\alpha I_E$ and the reverse saturation current of the collector base diode. A small fraction of the emitter current $-I_E(1-\alpha)$ flows out of the base terminal. The reverse saturation current of the collector base diode I_{C0} is opposite in direction to $I_E(1-\alpha)$ and it flows into the base. The base current can now be written as
$$I_B = -I_E(1-\alpha)+I_{C0} \tag{3.24}$$
when the base is open $I_B = 0$
$$I_E(1-a) = I_{C0}$$
$$I_E = I_{CE0} = \frac{I_{C0}}{1-\alpha} \tag{3.25}$$
and
$$I_{CE0} = -(\beta+1)\, I_{C0} \tag{3 26}$$

Thus the reverse saturation current I_{CE0} when the base is open is β times the saturation current of the reverse biased diode.

The following points should be borne in mind before discussing the characteristics of the transistor in the common emitter configuration.

1. The reverse saturation current I_{CE0} is many times larger than I_{C0}.
2. The common emitter current gain β is very large and small variations in α give rise to large variations in β.

Input characteristic

The input characteristic of a transistor in the common emitter configuration is the curve showing the relation between the base current and the base emitter voltage for different values of the collector voltage. This characteristic is different from that of a forward biased diode for the following reasons:

1. The base current is the difference between the emitter current and the collector current.
2. The applied base emitter voltage is larger than the voltage appearing across the base emitter junction due to the potential

drop across the base spreading resistance. If V_{EB} is the applied voltage the voltage appearing across the base emitter junction

$$V_E = V_{EB} - r_b I_B \tag{3.27}$$

The emitter current

$$I_E \approx I_0 \exp \frac{qV_E}{kT}$$

and

$$I_C \approx -\alpha I_0 \exp \frac{qV_E}{kT}$$

$$I_B = -I_E - I_C$$

$$= -I_0 (1-\alpha) \exp \frac{qV_E}{kT} \tag{3.28}$$

The minus sign indicates that the current flows out of the base.

The input characteristic curves in common base arrangement become steeper with an increase in the collector base voltage as reasoned before. In common emitter configuration the input curves become less steeper as V_{CE} is increased. The difference between the change in the slope of the input curves with increase in collector voltage is understandable if we note that in the common base arrangement the input curve shows the relation between I_E and V_{EB} while in the common emitter configuration the curve gives the relation between I_B and V_{EB} · I_B being the difference between I_E and I_C. When the common base current gain increases due to narrowing of the base region consequent on the increase in collector voltage the base current decreases. This is the reason why the slopes of the input curves in the common emitter configuration decrease with the increase in collector voltage.

Another important feature of the input characteristic is that the base current starts flowing out only after the base emitter voltage reaches a certain value which depends on whether the collector emitter voltage is zero or has a negative value. When the collector and emitter are at the same potential $V_{CE}=0$ the base emitter and the collector base junction voltages are very small and the large electron current flowing out of the base gives rise to a negative base current (base current flowing into the transistor) the potential rise in the base spreading resistance causes further reduction in the junction voltage. However, when the forward voltage reaches a certain value, the emitter base junction is forward biased and the density of holes on the base side of the base emitter junction reaches a larger value than the density of holes on the base side of the collector base junction. Due to the concentration gradient and due to the geometry of the junctions the hole current flows from emitter to collector and normal transistor action occurs.

When V_{CE} is made negative the collector junction acquires a negative voltage and the reverse saturation current I_{C0} flows into the base for small values of V_{BE} and due to the reverse current there is a potential rise in the base spreading resistance which in turn causes a reduction in the junction voltage. This continues till the base emitter voltage reaches a value making the base current flowing out of the transistor more than I_{C0} and the normal transistor action occurs.

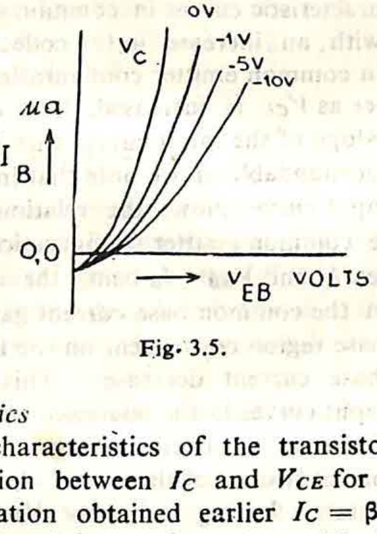


Fig. 3.5.

Output characteristics

The output characteristics of the transistor are the curves showing the relation between I_C and V_{CE} for different values of I_B. From the relation obtained earlier $I_C = \beta I_B$. The collector current should depend only on I_B but with the increase in V_{CE} there is an increase in the reverse voltage on the collector junction and the depletion layer moves into the base region of the transistor reducing the electrical base width, resulting with an increase in α. Even when α increases by a small value, β shows a marked increase. The output characteristics show an upward slope with increase in the output voltage V_{CE}.

$I_B=0$ corresponds to the condition $-I_C=I_E$, and as discussed earlier, the collector current has a value many times (β times) the value of reverse saturation current of a junction.

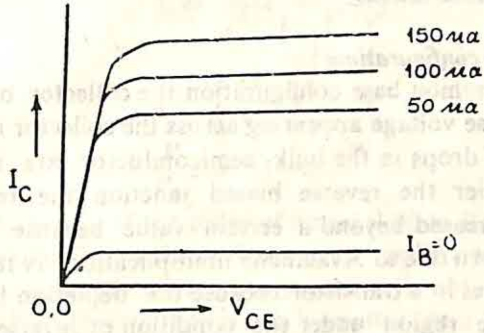


Fig. 3.6.

Transfer characteristic

The transfer characteristic is the curve showing the relation between the base current and collector current for different values of V_{CE}. The shape of the curve is obtained from the relation $I_C = \beta I_B$. As β depends on the emitter efficiency, transport factor, the curves are not straight lines. For small currents the emitter efficiency increases with the current and the curve has a slight upward bend. For currents in the mid range the curves are fairly linear indicating a constant value for β. At high currents the curves drops due to the fall in α. As the value of α increases with the negative voltages on the collector the gradient of the curve in the mid current region depends on the voltage between the emitter and collector.

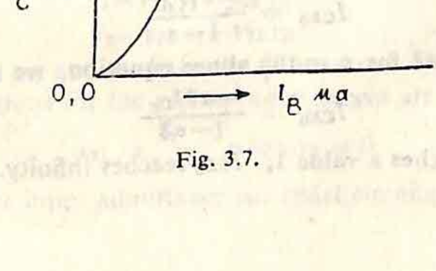


Fig. 3.7.

3.3 Voltage Limitations

Common base configuration

In the common base configuration the collector base voltage is also the reverse voltage appearing across the collector base junction, as the voltage drops in the bulk semiconductor are neglected. As discussed under the reverse biased junction the reverse voltage cannot be increased beyond a certain value because the junction may break down due to Avalanche multiplication. A more stringent limitation arises in a transistor because the depletion layer spreads into the base region under the condition of reverse bias. As the base region is very thin, much before the Avalanche breakdown occurs, the depletion layer may extend throughout the base region to touch the depletion layer on the emitter side. When this occurs, carriers are injected direct into the collector and the transistor fails. This failure is called punch through. The punch through voltage is less than the Avalanche breakdown voltage. The maximum voltage that can be applied is determined by the punch through voltage.

Common emitter configuration

The maximum voltage that can be applied in the common emitter configuration is much less than the voltage for the common base configuration as discussed below.

When discussing the collector multiplication factor (δ) we have seen that the factor δ is unity under normal operating conditions but increases with the reverse voltage according to the formula

$$\delta = \frac{1}{\left(1 - \frac{V}{V_{br}}\right)^n} \tag{2.58}$$

where n has value between 3 and 6. If the collector multiplication factor is taken into consideration the current gain of a transistor has a value $\alpha\delta$. The reverse saturation current in the common emitter configuration according to equation (3.25).

$$I_{CE_0} = \frac{-\alpha I_{C0}}{1-\alpha}$$

Substituting $\alpha\delta$ for α in the above equation, we have

$$I_{CE_0} = \frac{-\alpha\delta I_{C0}}{1-\alpha\delta}$$

When $\alpha\delta$ reaches a value 1, I_{CE_0} reaches infinity.

$$1 = \frac{\alpha}{1 - \left(\frac{V}{V_{br}}\right)^n} \tag{3.29}$$

$$1 - \left(\frac{V}{V_{br}}\right)^n = \alpha$$

$$\frac{V}{V_{br}} = (1-\alpha)^{1/n} \tag{3.30}$$

As α is close to unity, of the order of about 0.98 to 0.99, $(1-\alpha)$ has a very small value. The voltage that can be applied to the collector base junction without making I_{CB_0} excessive is obtained from the expression (3.30). This voltage is much smaller than the Avalanche breakdown voltage or punch through voltage and is called the sustaining voltage. The maximum voltage that can be applied in the common emitter configuration is limited by the sustaining voltage and is very much smaller than the breakdown voltage.

3.4.1 *A.C. behaviour of a junction transistor*

A transistor is used in various electronic circuits as an active element. When a transistor is used in an electronic circuit the base is given a steady voltage relative to the emitter so that the emitter base junction is forward biased, the collector is given a steady voltage relative to the base so that the collector base junction is reverse biased. Thus the transistor is brought into an active state emitter emitting and collector collecting. The steady voltages applied are called bias voltages and the transistor is said to be normally biased. The alternating voltages superimposed on the steady voltages give rise to alternating currents. Relations have to be established between the various alternating currents and voltages. Often the transistor is considered as a black box giving rise to definite changes in voltages and currents. If v_2 and i_2 are the output alternating voltage and current they are related to the input alternating voltage and current v_1, i_1 by two equations

$$i_1 = y_{11}\, v_1 + y_{12}\, v_2 \tag{3.31}$$

$$i_2 = y_{21}\, v_1 + y_{22}\, v_2 \tag{3.32}$$

where y_{11}, y_{12}, y_{21} and y_{22} have the dimensions of admittances. The defining equations for the various admittances are

$$y_{11} = \frac{i_1}{v_1} \quad \text{when } v_2 = 0 \tag{3.33}$$

y_{11} is called the input admittance for short circuited output.

$$y_{12} = \frac{i_1}{v_2} \quad \text{when } v_1 = 0 \tag{3.34}$$

y_{12} is called the reverse transfer admittance for short circuited input.

$$y_{21} = \frac{i_2}{v_1} \quad \text{when } v_2 = 0 \tag{3.35}$$

y_{21} is called the forward transfer admittance for short circuited output.

$$y_{22} = \frac{i_2}{v_2} \quad \text{when } v_1 = 0 \tag{3.36}$$

y_{22} is called the output admittance for short circuited input.

3.4.2 *Stored base charge*

The A.C. behaviour of a transistor can be understood and analysed if the various admittances are known. The problem now reduces to one of the calculation of the various admittances. The admittances can be found by writing down the diffusion equations with time dependent boundary conditions and solving them. But this procedure is as such involved and is more complicated by the geometry of the actual transistor.

The admittances can also be calculated from the fact that the alternating currents originate from charges that vary periodically. This method of approach is called the charge control method and we will use this method to evaluate the various admittances at low frequencies. The charge in the base region of a transistor can easily be calculated and from it the changes i_1 and i_2 be found. We will first calculate the charge in the base region. For convenience a *PNP* transistor is considered and the charge in base region is due to holes diffusing into the base from the emitter. Excess charge due to holes migrating from the *P* region to *N* region was calculated in Chapter 2 on the assumption that the *N* region extends over a sufficient distance so that the excess hole density reduces to zero exponentially. This assumption has to be dropped in a transistor because the hole density drops to zero at the edge of the depletion layer existing at the collector junction. The hole density in the base region is given by the expression

$$p_x = C_1 \exp \frac{x}{L_p} + C_2 \exp \frac{-x}{L_p} \tag{3.5}$$

As the base width is very small compared to the diffusion length of the holes in the base region and as the aim of the derivation is to evaluate the value of the base charge from the values of carrier densities at various points in the base the exponentials can be expanded and approximated to the first term. When this approximation is made, expression (3.5) is reduced to

$$p_x = C_1+C_2+\frac{x}{L_p}(C_1-C_2) \tag{3.37}$$

The boundary conditions for holes in the base region are

$$\left.\begin{array}{ll} \text{at} & x = 0;\ p_x = p_1 \\ \text{at} & x = W;\ p_x = 0 \end{array}\right\} \tag{3.38}$$

With these boundary conditions the constants C_1 and C_2 have values

$$\left.\begin{array}{l} C_1 = \dfrac{p_1}{2}\left(1-\dfrac{L_p}{W}\right) \\ C_2 = \dfrac{p_1}{2}\left(1+\dfrac{}{W}\right) \end{array}\right\} \tag{3.39}$$

The carrier density at a distance x from the edge of the depletion layer on the emitter side is given by the expression

$$p_x = p_1\left(1-\frac{x}{W}\right) \tag{3.40}$$

The charge due to holes in the base region

$$Q = Aq\int_0^W p_x dx = \frac{A \cdot qp_1W}{2} \tag{3.41}$$

The hole component of the emitter current at the edge of the depletion layer

$$I_p = -q \cdot AD_p \left.\frac{dp}{dx}\right|_{x=0} \tag{3.42}$$

$$= \frac{AqD_pp_1}{W} \tag{3.43}$$

Substituting value of p_1 from equation (3.43) into (3.41), we have

$$Qs = \frac{W^2}{2D_p} \cdot I_p \tag{3.44}$$

If the emitter efficiency is close to unity, $I_p = I_E$ and

$$Qs = \frac{W^2}{2D_p} \cdot I_E \tag{3.45}$$

when an alternating current flows through the transistor the charge in the base region is composed of a charge due to the steady current and a charge due to the alternating component of the current. The charge due to the alternating component is periodic. If the alternating voltage is sinusoidal with an angular frequency ω the base charge due to the alternating component can be written as

$$q_S = Q_S \exp j\omega t \tag{3.46}$$

where Q_S is the maximum charge. After determining the charge set up in the base region one should know the manner in which the charge is distributed. Part of this charge is associated with the junction and part of the charge flows away. To evaluate the charge stored on the junction we will calculate the admittance of the junction.

3.4.3 *Admittance of a forward biased junction*

Let a small alternating voltage represented by $v = v_1 \exp j\omega t$ be superimposed on a steady forward voltage V existing across the base emitter junction.. The hole density on the base side of the base emitter junction is given by the expression

$$p = p_0 \exp \frac{q}{kT}(V + v_1 \exp j\omega t) \tag{3.47}$$

From this expression we note that the maximum value of the hole density on the base side

$$p_1' = p_0 \exp \frac{q}{kT}(V + v_1)$$

as v_1 is very small, $\exp \frac{qv_1}{kT}$ can be written as $1 + \frac{qv_1}{kT}$ and

$$p_1' = p_0 \exp \frac{qV}{kT}\left(1 + \frac{qv_1}{kT}\right) \tag{3.48}$$

But $p_0 \exp \frac{qV}{kT}$ is the hole density on the base side due to the forward voltage V_1 represented by p_1

$$p_1' = p_1\left(1 + \frac{qv_1}{kT}\right)$$

$$\tilde{p}_1 = p_1 \frac{qv_1}{kT} \tag{3.49}$$

$\tilde{p}_1$ is the maximum change in the hole density occurring at the edge of the depletion layer on the base side. This change in hole density occurs throughout the base region and the expression for the maximum change in hole density at a distance x from the junction is given by the expression

$$\tilde{p} = C_1 \exp\,(x\tau) + C_2 \exp\,(-x\tau) \tag{3.50}$$

where C_1 and C_2 are constants to be evaluated from the boundary conditions, and

$$\tau = \frac{\sqrt{1+j\omega\tau_p}}{L_p} \tag{3.51}$$

(The above expression for the carrier profile was obtained in Chapter 2.) The boundary conditions for the base are

$$\text{when} \qquad x = 0 \quad \tilde{p} = \tilde{p}_1$$

$$x = W \quad \tilde{p} = 0$$

When these boundary conditions are applied to expression (3.50), we find that

$$C_1 = \frac{-\tilde{p}_1 \exp(-W\tau)}{2 \sinh W\tau} \tag{3.52}$$

$$C_2 = \frac{\tilde{p}_1 \exp\,(W\tau)}{2 \sinh W\tau}$$

Substituting the values of the constants C_1 and C_2 in equation (3.50) we obtain the expression giving the maximum change in carrier density with distance as

$$\tilde{p}_x = \frac{-\tilde{p}_1 \exp\,(-W\tau) \exp\,(x\tau) + \tilde{p}_1 \exp\,(W\tau) \exp\,(-x\tau)}{2 \sinh W\tau} \tag{3.53}$$

The alternating component of the hole current at the emitter junction can now be calculated from the formula

$$i = AqD_p \left.\frac{d\tilde{p}}{dx}\right|_{x=0} \tag{3.54}$$

$$i = Aq\,D_p\,\tau\,\tilde{p}_1 \coth W\,\tau \tag{3.55}$$

Substituting the value of $\tilde{p}_1$ from equation (3.49) we have

$$i = AqD_p\tau p_1 \frac{qv_1}{kT} \coth\; W\tau \tag{3.56}$$

From equation (3.8) we find that the steady hole current through the junction

$$I_p = \frac{AqD_p p_1}{L_p} \coth \frac{W}{L_p} \tag{3.8}$$

Substituting the value of p_1 from equation (3.8) into equation (3.56) we find

$$i = \frac{qv_1}{kT} \quad I_p \tau L_p \tanh \frac{W}{L_p} \coth W\tau \tag{3.57}$$

In the base region $L_p \gg W$ and $\tanh \frac{W}{L_p} \approx \frac{W}{L_p}$. Substituting the value of $L_p^2 = D_p\tau_p$ and writing $\frac{W}{L_p}$ for $\tanh \frac{W}{L_p}$ we have

$$i = \frac{qv_1}{kT} \cdot I_p \cdot L_p \frac{\sqrt{(1+j\omega\tau_p)}}{L_p}\frac{W}{L_p} \coth \sqrt{\frac{W^2}{L_p^2} + j\omega\frac{W^2}{D_p}} \tag{3.58}$$

$$= \frac{qv_1}{kT} I_p \sqrt{\frac{W^2}{L_p^2} + j\omega\frac{W^2}{D_p}} \coth \sqrt{\frac{W^2}{L_p^2} + j\omega\frac{W^2}{D_p}} \tag{3.59}$$

W^2/L_p^2 is very small and hence can be neglected in comparison with $j\omega\frac{W^2}{D_p}$ with the approximation

$$i = \frac{qv_1}{kT} I_p \sqrt{j\omega\frac{W^2}{D_p}} \coth \sqrt{j\omega\frac{W^2}{D_p}} \tag{3.60}$$

expanding $\coth \sqrt{j\omega\frac{W^2}{D_p}}$ and neglecting higher powers of W we can get the low frequency value for i

$$i = \frac{qv_1}{kT} \quad I_p \sqrt{j\omega\frac{W^2}{D_p}} \cdot \frac{1 + j\omega\frac{W^2}{2D_p}}{\sqrt{j\omega\frac{W^2}{D_p}} + j\omega\frac{W^2}{6D_p}\sqrt{j\omega\frac{W^2}{D_p}}} \tag{3.61}$$

$$= \frac{qv_1}{kT} I_p \quad \left(1 + j\omega\frac{W^2}{3D_p}\right) \tag{3.62}$$

From this expression it is clear that the current has a reactive component given by the expression

$$i_{\text{reactive}} = \frac{qv_1}{kT} \quad I_p \quad \frac{j\omega W^2}{3D_p} \tag{3.63}$$

The reactive part of the current is associated with a capacitance called the diffusion capacitance and the diffusion capacitance in case of the base emitter junction

$$C_{\text{diff}} = \frac{q}{kT} \quad I_p \quad \frac{W^2}{3D_p} \tag{3.64}$$

As the emitter efficiency is close to unity I_p can be replaced by I_E and

$$C_{\text{diff}} = \frac{q}{kT}\ I_E \ \cdot \ \frac{W^2}{3D_p} \tag{3.65}$$

as $\frac{q}{kT} I_E = \frac{dI}{dV}$ for a forward biased junction, substituting the value for $\frac{q}{kT} I_E$ we have

$$C_{\text{diff}} = \frac{W^2}{3D_p} \ \ \frac{dI}{dV} \tag{3.66}$$

and the charge on the diffusion capacitance is given by the expression ·

$$Q_{\text{diff}} = \frac{W^2}{3D_p}\, dI \tag{3.67}$$

From the above expression it is clear that when the current increases by the amount dI the charge on the base side of the emitter base junction increases by $\frac{W^2}{3D_p} \cdot dI$. From the expression.(3.45) we find that when the current increases by dI the increase in the stored base charge is $\frac{W^2}{2D_p}\, dI$. From these considerations we can conclude that two-thirds of the excess base charge resides on the junction capacitance.

3.4.4 *Evaluation of the admittance* y_{11} aad y_{21}

The input admittance is the ratio of the input current to the input voltage when the output voltage is zero.

When a sinusoidal voltage represented by $v = v_1 \exp j\omega t$ is applied across the input terminals the input current has two components. A real component which flows through the junction and a reactive component associated with the junction capacitance.

The real component $= i_{1\,(\text{real})} = \frac{v_1}{r_e}$ (3.68)

where r_e is the incremental resistance of the forward biased junction.

$$r_e = \frac{q}{kT}\ I_E$$

The charge on the diffusion capacitance is given by

$$q_e = \frac{q}{kT}\ I_p\, \frac{W^2}{3D_p} \cdot v_1 \exp j\omega t \tag{3.69}$$

The current associated with this junction capacitance

$$\frac{dq_e}{dt} = \frac{q}{kT} I_p \frac{W^2}{3D_p} j\omega v_1 \exp j\omega t \tag{3.70}$$

dropping the subscript exp $j\omega t$ as the current is understood to be sinusoidal, the reactive part of the input current

$$i_{(\text{react})} = \frac{q\,I_p}{kT} \quad \frac{W^2}{3D_p} j\omega v_1 \tag{3.71}$$

Writing $\frac{I_p}{I_E} = \alpha$, the current gain, we have

$$i_{1(\text{react})} \quad \frac{q\alpha I_E}{kT} \quad \frac{j\omega W^2}{3D_p} v_1 \tag{3.72}$$

The input current

$$i_1 = v_1 \frac{qI_E}{kT}\left(1 + j\omega\alpha \frac{W^2}{3D_p}\right)$$

and

$$y_{11} = \frac{qI_E}{kT}\left(1 + j\omega\alpha \frac{W^2}{3D_p}\right) \tag{3.73}$$

The output current i_2 flowing under the influence of the input voltage can also be evaluated. The input voltage v_1 gives rise to a current $\frac{v_1 qI_E}{kT}$ in phase with the applied voltage and flowing across the emitter junction. The fraction of this current reaching the collector junction is $\frac{\alpha v_1 qI_E}{kT}$. This is the real component of i_2.

According to the sign convention the current is positive when it flows into the black box (transistor), as such the real component of i_2

$$i_{2\,(\text{real})} = \frac{-\alpha v_1 qI_E}{kT} \tag{3.74}$$

The charge on the diffusion capacitance is given by the expression (3.69), this charge is two-third of the charge due to holes migrating to the base side. Therefore the excess base charge on the base side that is not associated with the diffusion capacitance can be written as

$$q_s - q_e = \frac{q\,I_p}{kT} \quad \frac{W^2}{6D_p} \cdot v_1 \exp j\omega t \tag{3.75}$$

The time dependence of this charge gives rise to a current that flows through the base region from the collector to the emitter side as such the imaginary component of i_2

$$\begin{aligned} i_{2(\text{imag})} &= \frac{q\,I_p}{kT} j\omega \frac{W^2}{6D_p} v_1 \\ &= \frac{qI_E}{kT} \alpha j\omega \frac{W^2}{6D_p} v_1 \end{aligned} \tag{3.76}$$

The output current i_2

$$i_2 = i_{2(\text{real})} + i_{2\,(\text{imag})}$$
$$= \frac{q\alpha I_E}{kT}\left(-1+j\omega\,\frac{W^2}{6D_p}\right)v_1 \qquad (3.77)$$

Therefore the forward transfer admittance

$$y_{21} = \frac{q\alpha I_E}{kT}\left(-1+j\omega\,\frac{W^2}{6D_p}\right) \qquad (3.78)$$

3.4.5 *Admittances y_{12} and y_{22}*

The output admittance y_{22} and the reverse transfer admittance y_{12} arise out of an effect first described by Early and known as Early effect. The collector junction of a transistor is reverse biased and the depletion layer spreads into the base region. The extent to which the depletion layer extends into the base region depends on the voltage across the base collector junction. The spreading of the depletion layer into the base region reduces the electrical base width and this in turn produces changes in emitter and collector currents.

The hole component of the emitter current at the edge of the emitter base depletion layer is given by the expression

$$I_p = \frac{AqD_p}{L_p}\cdot p_1 \coth\frac{W}{L_p} \qquad (3.8)$$

As $W \gg L_p$, tanh $\frac{W}{L_p}$ can be written as $\frac{W}{L_p}$

and

$$I_p = \frac{AqD_p\,p_1}{W} \qquad (3.79)$$

We can write

$$\frac{\partial I_p}{\partial V_C} = \frac{\partial I_p}{\partial W}\cdot\frac{\partial W}{\partial V_C} \qquad (3.80)$$

Substituting the value of $\partial I_p/\partial W$ from expression (3.79) we have

$$\frac{\partial I_p}{\partial V_C} = -\,\frac{AqD_{p1}p}{W^2}\,\frac{\partial W}{\partial V_C}$$
$$= -\,\frac{I_p}{W}\,\frac{\partial W}{\partial V_C} \qquad (3.81)$$

If the output voltage is sinusoidal the hole current variations are also sinusoidal. The real part of output current

$$i_{2\,(\text{real})} = \frac{-\alpha I_E}{W}\cdot\frac{\partial W}{\partial V_C} \qquad (3.82)$$

The sign change is because of the convention.

To evaluate the imaginary part of i_2 we should investigate the changes in the base charge associated with the movement of the depletion layer. If I_p is the hole current flowing when the base width is W we have from equation (3.79)

$$I_p = \frac{AqD_p p_1}{W}$$

and the associated base charge from equation (3.44)

$$Q = \frac{W^2}{2D_p} \quad \frac{AqD_p p_1}{W}$$

If the depletion layer shifts by an amount dW towards the emitter side the new hole current (I'_p)

$$I_p' = \frac{Aq\, D_p\, p_1}{W-dW} \tag{3.83}$$

and the base charge due to holes

$$Q' = \frac{(W-dW)^2}{2D_p} \cdot \frac{Aq\, D_p\, p_1}{(W-dW)} \tag{3.84}$$

The change in the base charge due to the movement of the depletion layer is

$$Q-Q' = \frac{Aq\, p_1}{2}\, dW \tag{3.85}$$

As proved earlier two-thirds of this charge is associated with the collector junction and the charge dQ_C associated with the collector junction due to the depletion layer is

$$dQ_C = \frac{1}{3}\, Aq\, p_1\, dW \tag{3.86}$$

Writing

$$\frac{\partial Q_C}{\partial V_C} = \frac{\partial Q_C}{\partial W} \; \frac{\partial W}{\partial V_C}$$

we have

$$\frac{\partial Q_C}{\partial V_C} = \frac{1}{3}\, Aq\, p_1\, \frac{\partial W}{\partial V_C} \tag{3.87}$$

When the voltage v_2 is sinusoidal the charge associated with the collector junction also varies sinusoidally and the charge can be written as

$$q = \frac{1}{3}\, Aq\, p_1\, \frac{\partial W}{\partial V_C}\, v_2 \exp j\omega t \tag{3.88}$$

The imaginary component of i_2 is

$$i_{2\,(\text{imag})} = \frac{1}{3}\, Aq\, p_1\, j\omega v_2 \cdot \frac{\partial W}{\partial V_C} \tag{3.89}$$

Substituting the value of $Aq\, p_1$ from equation (3.79) we have

$$i_{2\,(\text{imag})} = j\omega \frac{W}{3D_p}\, \alpha I_E v_2 \cdot \frac{\partial W}{\partial V_C} \tag{3.90}$$

The output current i_2 is therefore

$$i_2 = \frac{\alpha I_E}{W}\,\frac{\partial W}{\partial V_C}\left(1+j\omega\,\frac{W^2}{3D_p}\right)v_2 \qquad (3.91)$$

and the output admittance

$$y_{22} = \frac{\alpha I_E}{W}\,\frac{\partial W}{\partial V_C}\left(1+j\omega\frac{W^2}{3D_p}\right) \qquad (3.92)$$

Reverse transfer admittance y_{12}

The real part of the current i_2 flowing into base from collector side also flows through the emitter junction as such the real part i_2 when $v_1 = 0$ and a voltage v_2 applied across output is

$$= \frac{-\alpha I_E}{W}\,\frac{\partial W}{\partial V_C} \qquad (3.93)$$

In addition to this there is a transfer of charge through the base region as only two-thirds of excess base charge is associated with the collector junction. The charge flowing from emitter side to collector side is the difference between the charges calculated in expressions (3.85) and (3.86).

The magnitude of this charge from expression (3.88) is

$$= \frac{1}{6}\,Aq\,p_1\,\frac{\partial W}{\partial V_C}\,v_2 \exp j\omega t \qquad (3.94)$$

The current associated with this charge is

$$i_{1\,(\text{imag})} = \frac{\partial W}{\partial V_C}\,j\omega\,\frac{W}{6D_p}\,\alpha I_E\,v_2 \qquad (3.95)$$

The input current i_1 when $v_1 = 0$ is now obtained as

$$i_1 = \frac{\alpha I_E}{W}\,\frac{\partial W}{\partial V_C}\left(-1+j\omega\,\frac{W^2}{6D_p}\right)v_2 \qquad (3.96)$$

and the reverse transfer admittance

$$\frac{i_1}{v_2} = y_{12} = \frac{\alpha I_E}{W}\,\frac{\partial W}{\partial V_C}\left(-1+j\omega\,\frac{W^2}{6D_p}\right) \qquad (3.97)$$

3.4.6 *Transit time of holes through the base region of a transistor*

If I_p is the hole current flowing from emitter to base we have from the diffusion equation

$$I_p = Aq\frac{dp'}{dx}$$

where p' is the excess density of holes at the point. The charge crossing the section in unit time is qAp' and the velocity of holes at the section is

$$v_p = \frac{I_p}{qAp}$$

The time taken by the holes to travel through the base or the transit time

$$\tau_d = \int_0^W \frac{dx}{v_p}$$

$$= \frac{1}{I_p}\int_0^W qAp'\,dx = \frac{Q_s}{I_p}$$

but

$$Q_s = \frac{W^2}{2D_p}\,I_p$$

Combining the above equation we find

$$\tau_d = \frac{W^2}{2D_p} \qquad (3.98)$$

The various admittances calculated earlier may be split into real and imaginary parts and expressed in the terms of the transit time.

Writing

$$\frac{qI_E}{kT} = g_1$$

$$\frac{\alpha I_E}{W}\,\frac{\partial W}{\partial V_C} = g_2$$

$$\frac{W^2}{2D_p} = \tau_d$$

We have

$$\left.\begin{aligned} y_{11} &= g_1\left(1+\frac{2}{3}\,\alpha\, j\omega\tau_d\right)\\ y_{12} &= g_2\left(-1+\frac{1}{3}\,j\omega\tau_d\right)\\ y_{21} &= -g_1\,\alpha\left(1-\frac{1}{3}\,j\omega\tau_d\right)\\ y_{22} &= g_2\left(1+\frac{2}{3}\,j\omega\tau_d\right)\end{aligned}\right\} \qquad (3.99)$$

3.4.7 *Equivalent circuit of a transistor*

The equivalent circuit of a transistor is constructed from the defining equations of the four pole parameters. The first equation $i_1 = y_{11}v_1 + y_{12}v_2$ suggests that the input current i_1 has two components $y_{11}v_1$ and $y_{12}v_2$. The component $y_{11}v_1$ is due to the input voltage and is the result of an admittance y_{11} across the input terminals. From the expression for the input admittance evaluated earlier we find that it is due to a parallel combination of a resistance of value $1/g_1$ and a capacitance of value $\frac{2}{3}g_1\tau_d$. The component $y_{12}v_2$ of the input current is due to a current generator on the input side generating a current $y_{12}v_2$.

The second equation $i_2=y_{21}\,v_1+y_{22}v_2$ suggests that the output current i_2 is due to a current generator $y_{21}v_1$ and an admittance y_{22} existing across the output terminals. The output admittance is again a parallel combination of a resistance $1/g_2$ and a capacitance $\frac{2}{3}\,g_2\tau_d$.

From the above considerations we construct an equivalent circuit as under.

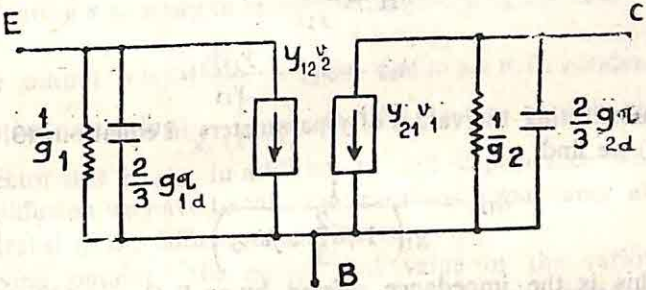


Fig. 3.8.

In drawing the above equivalent circuit we have not taken into account the transition and lead capacitances existing across the base emitter and base collector junctions. These capacitances appear in parallel to the diffusion capacitances and the capacitances in the above equivalent circuit should be modified. The base spreading resistance r_b appears in series with the base lead and it should also be incorporated.

3.4.8 *h-parameters*

The *y* parameters evaluated earlier are not convenient to use in circuit analysis and another set of parameters called *h*-parameters are often used for transistor circuit analysis. The *h*-parameters are defined by the following set of equations.

$$\left.\begin{aligned} v_1 &= h_{11}i_1+h_{12}v_2 \\ i_2 &= h_{21}i_1+h_{22}v_2 \end{aligned}\right\} \tag{3.100}$$

h_{11} is the input impedance when the output is shorted; h_{12}—reverse transfer voltage ratio for an open circuited input; h_{21}—forward current transfer ratio for a short circuited output and h_{22}—output admittance for an open circuited input.

The relationships between the h-parameters and y-parameters can easily be obtained from equation (3.100) and equations (3.31) and (3.32). The relationships are

$$h_{11}=\frac{1}{y_{11}} \tag{3.101}$$

$$h_{12}=\frac{y_{12}}{y_{11}} \tag{3.102}$$

$$h_{21}=\frac{y_{21}}{y_{11}} \tag{3.103}$$

$$h_{22}=y_{22}-\frac{y_{12}y_{21}}{y_{11}} \tag{3.104}$$

Substituting the values of y-parameters in equations (3.101) to (3.104) we find

$$h_{11}=\frac{1}{g_1\left(1+\frac{2}{3}\,\alpha j\omega\tau_d\right)} \tag{3.105}$$

This is the impedance offered by an R.C. combination, the resistance having a value $1/g_1$ and the capacitance of value $\frac{2}{3}g_1\alpha\tau_d$; this is again the diffusion capacitance of the forward biased junction. In addition to this capacitance we should also take the transition capacitance of the emitter base junction.

$$h_{12}\doteq\frac{-g_2\left(-1+\frac{1}{3}\,j\omega\tau_d\right)}{g_1\left(1+\frac{2}{3}\,\alpha j\omega\tau_d\right)} \tag{3.106}$$

Neglecting the terms containing ω^2 and approximating α to 1 we finally have

$$h_{12}\approx\frac{g_2}{g_1}\left(1-j\omega\tau_d\right) \tag{3.107}$$

is the voltage generator in the input circuit which opposes the applied input voltage. This voltage is of the same phase as the output voltage at lower frequencies but there is a slight phase change introduced at higher frequencies

$$h_{21}=\frac{y_{21}}{y_{11}}$$

$$= \frac{-g_1 \alpha (1-\frac{1}{3} j\omega\tau_d)}{g_1 (1+\frac{2}{3} \alpha j\omega\tau_d)} \tag{3.108}$$

Neglecting terms involving ω^2 and higher powers and taking $\alpha=1$ in terms multiplying $j\omega\tau_d$

$$h_{21}=\alpha (1-j\omega\tau_d) \tag{3.109}$$

h_{21}, the A.C. current gain, depends on the frequency. The current gain decreases and a phase change is introduced at higher frequencies. (The variation of current gain with frequency will be discussed in a later section.)

$$h_{22} = y_{22} - \frac{y_{12}\, y_{21}}{y_{11}}$$

$$= g_2 (1+\frac{2}{3} j\omega\tau_d) - \frac{g_1\, g_2 (1-\frac{1}{3} j\omega\tau_d)^2}{g_1 (1+\frac{2}{3} \alpha\, j\omega\tau_d)} \tag{3.110}$$

Neglecting the terms involving ω^2 and higher powers of ω and approximating α to unity in terms multiplying $j\omega\tau_d$ we have

$$h_{22}=g_2 (1-\alpha)+\frac{2}{3} j\omega\tau_d\, g_2 \tag{3.111}$$

The output admittance is again due to an R.C. combination the resistive part being $\frac{1}{g_2 (1-\alpha)}$ and the junction capacitance on the collector side $\frac{2}{3} \tau_d g_2$. In addition to this capacitance which is due to diffusion we have to take the transition capacitance also to be in parallel to the diffusion capacitance.

Having obtained the nature and value of the various h-parameters we can construct the equivalent circuit based on h-parameters.

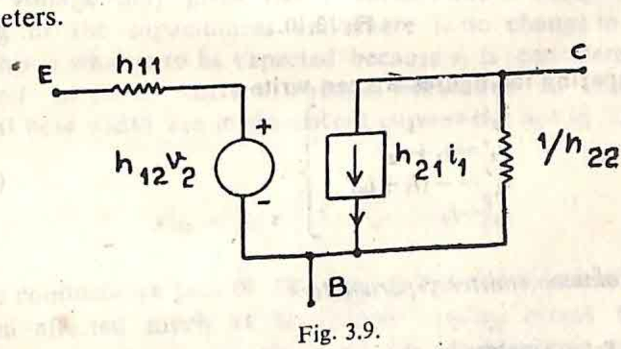
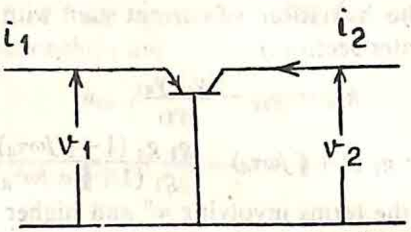


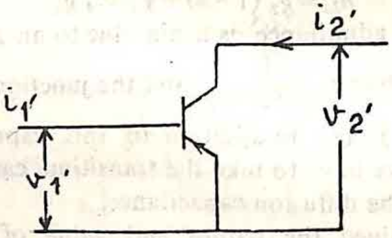
Fig. 3.9.

3.5.1 *Common emitter configuration*

Transistors are widely used in common emitter configuration. When a transistor is used in the common emitter configuration the input voltage is across the base and emitter while the output is across the collector and emitter. The input current is the current flowing into the base while the output current is the current flowing into the collector. The relationship between voltages and currents in the common base and common emitter configurations become clear when we draw the transistor in the two configurations and find the relation between voltages and currents.



COMMON BASE CONFIGURATION



COMMON EMITTER

Fig. 3.10.

Comparing the figures we can write

$$\left.\begin{aligned} v_1' &= -v_1 \\ v_2' &= v_1 + v_2 \\ i_1' &= -(i_1 + i_2) \\ i_2' &= i_2 \end{aligned}\right\} \qquad (3.112)$$

3.5.2 *Common emitter y-parameters*

The *y*-parameters in the common emitter configuration can easily be obtained from the *y*-parameters in common base configuration by substituting the values of common emitter voltages and currents in the defining equations of *y*-parameters given in equations (3.31) and (3.32).

The common emitter y-parameters are

$$\left.\begin{aligned} y_{11}' &= y_{11}+y_{12}+y_{21}+y_{22} \\ y_{12}' &= -(y_{12}+y_{22}) \\ y_{21}' &= -(y_{21}+y_{22}) \\ y_{22}' &= y_{22} \end{aligned}\right\} \qquad (3.113)$$

Substituting the values of the common base y-parameters in equations (3.113) and neglecting imaginary terms involving g_2 in comparison with the imaginary terms involving g_1 and approximating α to 1 in the imaginary terms we obtain

$$y_{11}' = g_1 \{(1-\alpha)+j\alpha\omega\tau_d\} \qquad (3.114)$$

From this expression it is obvious that the input admittance has decreased considerably. The conductance part of the admittance has decreased by $(1-\alpha)$ times. This is obvious from the fact that the input current is the base current in the common emitter configuration which is the difference of the emitter and collector currents. The susceptance part of the admittance has increased to indicate that the entire base charge effect can be associated with the base emitter junction

$$y_{12}' = -g_2\, j\omega\tau_d \qquad (3.115)$$

The absence of the real part in y'_{12} indicates that the movement of the depletion layer boundary due to variation in collector emitter voltage only gives rise to current due to charging or discharging of the capacitances and there is no change in the real part. This is what is to be expected because v_1 is considered to be zero and whatever current changes occur due to changes in electrical base width are in the output current but not in the input current

$$y'_{21} = g_1\, \alpha \left\{ 1-\frac{1}{3}\, j\omega\tau_d \right\} \qquad (3.116)$$

The conductance part of the forward transfer admittance has not been affected much as the voltage applied across the base emitter junction is almost the same in the common base and common emitter configurations

$$y_{22}' = g_2 \left(1+\frac{2}{3}\, j\omega\tau_d \right) \qquad (3.117)$$

The output admittance has the same value.

3.5.3 *Common emitter h-parameters*

After obtaining the admittances in the common emitter configuration we can easily obtain the *h*-parameters in the common emitter configuration

$$h'_{11} = \frac{1}{y_{11}'}, = \frac{1}{g_1\,(1-\alpha)+g_1\,j\omega\tau_d} \tag{3.118}$$

This only shows that the input impedance of the transistor is due to a parallel combination of a conductance $g_1\,(1-\alpha)$ and the diffusion capacitance of value $g_1\tau_d$

$$h_{12}' = \frac{-y_{12}'}{y_{11}'} = \frac{g_2}{g_1\,(1-\alpha)}\,j\omega\tau_d \tag{3.119}$$

The reverse voltage transfer ratio has a large value because of the factor $(1-\alpha)$ in the denominator. This is because when the base is open for alternating currents, a variation in the collector emitter voltage directly varies the base emitter voltage also. In the common base arrangement the variation in the collector voltage directly affects the voltage across the collector base junction and the reverse voltage transfer is caused by changes in the electrical base width

$$h_{21}' = \frac{y_{21}}{y_{11}} \approx \frac{\alpha}{1-\alpha}\,\{1-j\omega\tau_d\} \approx \frac{\alpha}{1-\alpha} \tag{3.120}$$

The current gain in the common emitter configuration represented by h_{21}' is very large and a transistor in common emitter configuration gives a large current gain $\frac{\alpha}{1-\alpha}$, represented by β called the common emitter current gain. The capacitive part also increases considerably. The imaginary part is usually neglected in low frequency circuits

$$\begin{aligned} h_{22}' &= y_{22}' - \frac{y_{12}'\,y_{21}'}{y_{11}'} \\ &\approx g_2\left(1+\frac{\alpha}{1-\alpha}\,j\omega\tau_d\right) \end{aligned} \tag{3.121}$$

The output admittance in the common emitter configuration has a conductance much smaller than the conductance in the common base configuration by an order of $(1-\alpha)$ which is very small, but the susceptance part of the admittance has increased by β times its value in the common base configuration.

3.5.4 *Practical equivalent circuits*

Before we proceed to discuss the various practical equivalent circuits of transistors we should consider one useful parameter called mutual conductance. The mutual conductance is the ratio of the output current to the input voltage for a short circuited output.

When $v_2=0$

$$\left.\begin{aligned} v_1&=h_{11}i_1 \\ i_2&=h_{21}i_1 \end{aligned}\right\} \tag{3.122}$$

From the above equations we have

$$g_m=\frac{i_2}{v_1}=\frac{h_{21}}{h_{11}}=g_1\alpha+j\omega\beta g_1\tau_d \tag{3.123}$$

The first practical equivalent circuit is the hybrid π circuit and it is characterised by a single generator in contrast to two generators incorporated in the earlier circuits.

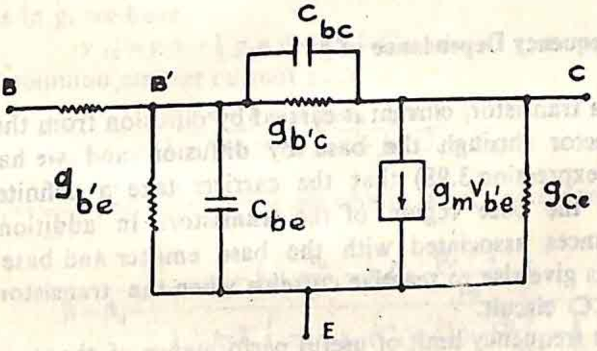


Fig. 3.11.

The above circuit is valid over a wide range of frequencies and it has an added advantage in that the base spreading resistance can be considered independently in series with the basic transistor structure connecting a point inside the base B' to the point B which is the external base terminal. The conductance $g_{b'e}$ and the capacitor C_{be} give rise to the input impedance. From equation (3.118)

$$g_{b'e}=g_1(1-\alpha) \quad \text{and} \quad C_{be}=g_1\tau_d \tag{3.123a}$$

$g_{b'c}$ is the conductance of the reverse biased junction and C_{bc} is the capacitance across the collector base junction.

The current generator in the output circuit generates a current givenby the experssion $g_m v_{b'e}$, where$v_{b'e}$ is the voltage appearing across the emitter base junction. The last component to be indentified in the equivalent circuit is the conductance g_{ce}. This is the conductance introduced to produce an effect similar to the reverse voltage transfer factor h_{12}. When the input is open ($i_1=0$), the voltage across the base emitter junction $v_{b'e}$ has a value equal to v_2h_{12}. As a consequence of this voltage, the current generator drives a current, $g_mh_{12}v_2$. The current flowing into the transistor with the input open is given by the expression v_2h_{22}, out of this current $g_mh_{12}v_2$ flows from the generator and the remaining current should flow through g_{ce}. (The effect of conductances $g_{b'c}$, and $g_{b'e}$ are neglected because the conductance of $g_{b'c}$ which is due to reverse biased junction, is very small.) The current through g_{ce}

$$\begin{aligned} v_2g_{ce} &= v_2h_{22}-g_mh_{12}v_2 \\ g_{ce} &= h_{22}-g_mh_{12} \\ &\approx g_2 \end{aligned} \tag{3.124}$$

3.6 Frequency Dependance of β

In a transistor, current is carried by diffusion from the emitter to collector through the base by diffusion and we have found earlier (expression 3.98) that the carriers take a definite time to traverse the base region of the transistor. In addition to this, capacitances associated with the base emitter and base collector junctions give rise to reactive currents when the transistor is used in an A.C. circuit.

The frequency limit of useful performance of the transistor is determined by the transit time through the base region and the time constants associated with the capacitive effects of the junctions. The common emitter configuration is widely used and the common emitter current gain β varies considerably with frequency. As such we will discuss the variation of common emitter current gain with frequency quantitatively.

The various admittances were evaluated for the transistor in the common base configuration and given by expressions (3.99). While calculating the admittances the transition capacitances were not taken into account. The transition capacitances and any other stray capacitances across the junction give rise to currents $j\omega C_1$ on the input side and $j\omega C_2$ on the output side, if C_1 and C_2 are taken as the capacitances connected between the base emitter and base collector to account for the transition and other stray capacitances. When the capacitances are taken into account the common base parameters are modified as

$$y_{11}=g_1+\tfrac{2}{3}\, g_1\alpha\, j\omega\tau_d+j\omega C_1$$
$$y_{22}=g_2+\tfrac{2}{3}\, g_2\, j\omega\tau_d+j\omega C_2$$
$$y_{12}=g_2(-1+\tfrac{1}{3}\, j\omega\tau_d)$$
$$y_{21}=g_1\alpha(1-\tfrac{1}{3}\, j\omega\tau_d)$$

The input admittance in the common emitter configuration

$$\begin{aligned} y'_{11}&=y_{11}+y_{12}+y_{21}+y_{22}\\ &=g_1\,(1-\alpha)+j\omega\tau_d\,(g_1\alpha+g_2)+j\omega\,(C_1+C_2) \end{aligned} \tag{3.125}$$

As g_2 is very small compared to g_1 it can be neglected in the imaginary terms. Neglecting g_2 in the imaginary terms we have

$$y'_{11}=g_1\,(1-\alpha)+j\omega\tau_d\, g_1\alpha+j\omega\,(C_1+C_2) \tag{3.126}$$

$$\begin{aligned} y'_{21}&=-(y_{21}+y_{22})\\ &=-(-g_1\alpha+\tfrac{1}{3}\, g_1\alpha\, j\omega\tau_d+g_2+\tfrac{2}{3}\, g_2\, j\omega\tau_d)-j\omega C_2 \end{aligned} \tag{3.127}$$

Again neglecting the terms involving g_2 in comparison with the terms in g_1 we have

$$y'_{21}=g_1\alpha-\tfrac{1}{3}\, g_1\alpha\, j\omega\tau_d-j\omega C_2 \tag{3.128}$$

The common emitter current gain

$$\frac{y_{21}'}{y_{11}'}=\frac{g_1\alpha-\frac{1}{3}\, g_1\alpha\, j\omega\tau_d-j\omega C_2}{g_1(1-\alpha)+j\omega\tau_d\, g_1\alpha+j\omega(C_1+C_2)}$$

writing $r_e=\dfrac{1}{g_1}$ and $\beta_0=\dfrac{\alpha}{1-\alpha}$ we obtain, after rearrangement

$$\beta=\beta_0\frac{1-j\omega\left(\dfrac{\tau_d}{3}-r_eC_2\,\dfrac{\beta_0+1}{\beta_0}\right)}{1+j\omega\beta_0\left(\tau_d+r_e\,(C_1+C_2)\,\dfrac{\beta_0+1}{\beta_0}\right)}$$

$$\left|\frac{\beta}{\beta_0}\right|=\left[\frac{1+\omega^2\left(\dfrac{\tau_d}{3}-r_eC_2\,\dfrac{\beta_0+1}{\beta_0}\right)^2}{1+\omega^2\beta_0^2\left(\tau_d+r_e(C_1+C_2)\,\dfrac{\beta_0+1}{\beta_0}\right)^2}\right]^{1/2} \tag{3.130}$$

The change in the value of β with frequency given by expression (3.130) indicates that β decreases with frequency. The contribution to the increase in β due to the numerator of the expression is negligible as compared to the decrease due to the increase in the value of the denominator. This is because of the fact that β_0 is very large. The cut off frequency is defined as the frequency at which the common emitter current gain reduces to $1/\sqrt{2}=(0.707)$

of its low frequency value. The cut off frequency can be evaluated by setting the denominator of expression (3.130) to 2 giving

$$\omega\,\beta_0\left(\tau_d+r_e(C_1+C_2)\,\frac{\beta_0+1}{\beta_0}\right)=1 \tag{3.131}$$

as β_0 is large $\qquad \dfrac{\beta_0+1}{\beta_0}\approx 1$

and the β cut off frequency is given by the expression

$$\frac{1}{\omega_{c\beta}}=\beta_0\left[\tau_d+r_e\,(C_1+C_2)\right]$$

$$\approx\beta_0\cdot\tau_d \tag{3.132}$$

As discussed earlier the reduction in β due to the increase in the denominator is much more than the increase due to the increase in the numerator and expression (3.130) can be written as

$$\frac{|\beta|}{\beta_0}=\left[\frac{1}{1+\omega^2\beta_0^{\,2}\,[\tau_d+r_e\,(C_1+C_2)]^2}\right]^{1/2} \tag{3.133}$$

If the capacitances shunting the junctions are neglected, we have

$$\frac{|\beta|}{\beta_0}=\frac{1}{(1+\omega^2\beta_0^{\,2}\tau_d^{\,2})^{1/2}}$$

At very high frequencies $\omega^2\beta_0^{\,2}\tau_d^{\,2}$ is very large compared to 1 and we can write

$$\frac{|\beta|}{\beta_0}=\frac{1}{\omega\beta_0\tau_d} \tag{3.134}$$

From the above expression it is clear that at high frequencies $|\beta|\,\omega$ is a constant. This constant is called the gain band width product for the common emitter current gain. The manufacturers specification of transistors often gives this product. From expression (3.134) it is also clear that the reciprocal of the gain-band width product is the transit time of the carriers through the base region of the transistor. At a frequency $1/\tau_d$ the common emitter current gain falls to unity.

3.7 High Current Effects in a Transistor

When a forward voltage is applied across the base emitter junction (*PNP* transistor) the hole density on the base side of the emitter junction increases to a large value. The electron density also increases to large value as the space charge neutrality has to be maintained at every point. The electron density changes from a large value, almost double the equilibrium value at the emitter junction to a small value at the collector junction. When the applied forward voltage makes the hole density equal to the electron density, this large change in concentration gives rise to an electric field within the base region.

The field within the base region can be calculated from the assumption that there is no electron current within the base region. The current flow equation for electrons in the base region

$$0 = q\mu_n nE + qD_n \frac{dn}{dx}$$

$$E = \frac{-D_n}{\mu_n n} \cdot \frac{dn}{dx} = \frac{-kT}{q} \frac{dn/dx}{n} \tag{3.135}$$

As space charge neutrality is maintained at every point $n=p$ and $\frac{dn}{dx} \quad \frac{dp}{dx}$. With these changes the field within the base region of a transistor

$$E = \frac{-kT}{q} \frac{dp}{dx} \frac{1}{p} \tag{3.136}$$

The hole current within the base region under the influence of the field

$$I_p = qAD_p \left(-\frac{dp}{dx} + \frac{qE}{kT} p \right) \tag{3.137}$$

Substituting the value of E we have

$$I_p = -2qAD_p \frac{dp}{dx}. \tag{3.138}$$

Comparing with the expression (3.42)

$$I_p = -qAD_p \frac{dp}{dx}$$

it is clear that the value of the diffusion constant has increased by a factor 2 in the presence of the field. The transit time of holes through the base region evaluated in expression (3.98) is therefore reduced by a factor 2. The reduction in transit time due to the field does not however improve the current gain of the transistor because the improvement in the gain due to the field is more than offset by the reduction in emitter efficiency.

When there is an increase in hole density on the base side of the emitter junction there is a corresponding increase in the electron density on the emitter side of the junction. Expression (3.9) giving the electron density on the emitter side is not a rigorous expression. The correct expression for the electron density on the emitter side of the junction is

$$n_{10}=n_{0E}\exp\frac{qv}{kT}\left(1+\frac{p_1}{n_{0B}}\right) \tag{3.139}$$

The factor p_1/n_{0B} was neglected earlier as the applied voltage is small and p_1 is much smaller than n_{0B} the equilibrium electron density in the base region. When the hole density on the base side of the junction approaches the electron density it is clear that the electron density on the emitter side increases to almost double the value it has under small forward voltages. This increase in electron density on the P side gives rise to a large electron current and consequently the emitter efficiency decreases to a small value. The overall effect on the current gain due to decrease in emitter efficiency and decrease in transit time is to decrease the current gain of the transistor. Added to this there is an increase in recombination due to large values of hole and electron densities.

3.8 Thermal Effects

When a transistor carries current an amount of power equal to the product of the collector voltage and current is dissipated at the collector junction (the region where the resistance is maximum). The power dissipated is converted to heat and the heat flows by conduction from the collector junction to the can of the transistor. As the transference of heat is by conduction the quantity of heat conducted away from the junction, which also represents the amount of power dissipated at the junction, is proportional to the temperature difference existing between the junction and the can. If T_j is the junction temperature and T_c the can temperature the power dissipated

$$V_C I_C=P\propto(T_j-T_c) \tag{3.140}$$

$$P=\frac{T_j-T_c}{K_{\mathrm{Th}}}$$

The constant of proportionality K_{Th} depends on the geometry of the device and the thermal conductivity of the material in the path of current flow. K_{Th} is called the thermal resistance of the transistor. The junction temperature of a germanium transistor is not to exceed about 50°C while the junction temperature of a silicon transistor may be about 200°C. From a knowledge of the can temperature and the permissible junction temperature the maximum power that can be carried by the transistor can be evaluated if the thermal resistance is known. Thermal resistance is provided with each transistor by the manufacturer.

3.9.1 *Charge control approach to transistor action*

The transistor theory presented earlier was based on diffusion equations. In the solution of the diffusion equations often an approximation is made stating that the applied alternating voltages are small. This approximation cannot be made when a transistor is used as a switch. The geometry of a transistor is of such a nature that plain diffusion equations cannot be written. Under these conditions a charge control approach to transistor theory is useful. We will present the charge control theory and use it to understand the working of a transistor as a switch. The hole density in the base region of a *PNP* transistor can be assumed to vary linearly from a maximum value on the emitter side of the base to a minimum or zero value on the collector side of the base. On this assumption the hole density at a distance x from the emitter base junction has been found as $p_x = p_1\left(1-\frac{x}{W}\right)$, where p_1 is the hole density on the emitter side of the base and W the electrical base width. The excess charge due to holes has also been evaluated as $Q_s = \frac{W^2}{2D_p} \cdot I_p$, where I_p is the hole current in the base. In the absence of recombination the hole current in the base is a constant and the collector current is equal to the hole current. This assumption was made earlier and the collector current of the *PNP* transistor was written down as $I_C = I_p = \frac{-Q_s}{\tau_d}$, where τ_d is the transit time of holes in the base region $\left(\tau_d = \frac{W^2}{2D_p}\right)$.

The recombination in the base region has to be taken into account in a rigorous analysis. The current loss due to recombination in the base region is proportional to the excess density of carriers, that is, to the charge stored and the recombination current can be written down as

$$I_r = \frac{Q_s}{\tau_r} \tag{3.141}$$

where τ_r is the lifetime of holes in the base region.

The hole current at the emitter base junction is the sum of the collector current and the current lost due to recombination and

$$I_p = I_C + I_r = Q_s\left(\frac{1}{\tau_d}+\frac{1}{\tau_r}\right)$$

$$= \frac{Q_s}{\tau_b}\left(\text{where } \frac{1}{\tau_b}=\frac{1}{\tau_d}+\frac{1}{\tau_r}\right) \qquad (3.142)$$

The charge associated with electrons on the emitter side can be found in terms of the stored charge in the base region and the emitter efficiency.

If τ_{nE} is the lifetime of electrons in the emitter region and Q_{nE} the charge associated with electrons moving across the base emitter junction from base to emitter, the electron current at the base emitter junction $I_{nE}=\frac{Q_{nE}}{\tau_{nE}}$ and the charge Q_{nE} is given by the expression

$$Q_{nE} = Q_s \cdot \frac{\tau_{nE}}{\tau_b} \cdot \frac{1-\gamma}{\gamma} \qquad (3.143)$$

The total charge moving across the base emitter junction

$$Q_E = Q_s + Q_{nE} = Q_s\left(1+\frac{\tau_{nE}}{\tau_b}\cdot\frac{1-\gamma}{\gamma}\right) \qquad (3.143)$$

Having calculated the total charge moving across the base emitter junction we can express the various currents in terms of charge Q_E. The collector current

$$I_C = \frac{-Q_s}{\tau_d} = \frac{-Q_E}{\tau_d\left\{1+\frac{\tau_{nE}}{\tau_b}\left(\frac{1-\gamma}{\gamma}\right)\right\}}$$

writing

$$\frac{1}{\tau_{n1}} = \frac{1}{\tau_d\left\{1+\frac{\tau_{nE}}{\tau_b}\left(\frac{1-\gamma}{\gamma}\right)\right\}} \qquad (3.144)$$

$$I_C = -\frac{Q_E}{\tau_{n1}}$$

The current due to recombination in the base region

$$I_r = \frac{Q_s}{\tau_r} = \frac{-Q_E}{\tau_r\left\{1+\frac{\tau_{nE}}{\tau_b}\left(\frac{1-\gamma}{\gamma}\right)\right\}} \qquad (3.145)$$

Electron current at the emitter junction

$$I_{nE} = \frac{Q_{nE}}{\tau_{nE}} = \frac{Q_E}{\tau_{nE}\left\{1+\frac{\tau_b}{\tau_{nE}}\left(\frac{\gamma}{1-\gamma}\right)\right\}} \qquad (3.146)$$

As the emitter current is the sum of the collector current, current due to recombination in the base region and the emitter current at the base emitter junction, it can be expressed in terms of the charge Q_E as

$$I_E = Q_E\left[\frac{1}{\tau_{n1}}+\frac{1}{\tau_r\left\{1+\frac{\tau_{nE}}{\tau_b}\cdot\frac{1-\gamma}{\gamma}\right\}}+\frac{1}{\tau_{nE}\left\{1+\frac{\tau_b}{\tau_{nE}}\cdot\frac{\gamma}{1-\gamma}\right\}}\right]$$

Writing

$$\frac{1}{\tau_r\left\{1+\frac{\tau_{nE}}{\tau_b}\cdot\frac{1-\gamma}{\gamma}\right\}}+\frac{1}{\tau_{nE}\left\{1+\frac{\tau_b}{\tau_{nE}}\cdot\frac{\gamma}{1-\gamma}\right\}}=\frac{1}{\tau_{nr}} \quad (3.147)$$

we have

$$I_E = \frac{Q_E}{\tau_{n1}}+\frac{Q_E}{\tau_{nr}}$$

$$=\frac{Q_E}{\tau_n} \quad (3.148)$$

where

$$\frac{1}{\tau_n}=\frac{1}{\tau_{n1}}+\frac{1}{\tau_{nr}} \quad (3.149)$$

Under steady state conditions

$$I_E-\frac{Q_E}{\tau_n}=0 \quad (3.150)$$

3.9.2 *Current gains of the transistor*

Under varying conditions the charge Q_E is related to the instantaneous emitter current I_E by the expression

$$\frac{dQ_E}{dt}=I_E-\frac{Q_E}{\tau_n} \quad (3.151)$$

If $Q_E(s)$ is the Laplace transform of the charge $Q_{(E)}$, we have from expression (3.151)

$$sF(s)=I_E(s)-\frac{F(s)}{\tau_n}$$

and

$$I_E(s)=F(s)\frac{(s\tau_n-1)}{\tau_n} \quad (3.152)$$

The Laplace transform of the collector current

$$I_C(s)=\frac{-F(s)}{\tau_{n1}} \quad (3.153)$$

The current gain of the transistor in Laplace notation

$$\alpha = \frac{\tau_n}{\tau_{n1}\,(1+s\tau_n)} \tag{3.154}$$

When the transistor is used A.C. circuits the current gain of the transistor is obtained from equation (3.154) by settlng $s=j\omega$. The current gain of the transistor at an angular frequency ω is therefore

$$\alpha = \frac{\tau_n}{\tau_{n1}(1+j\omega\tau_n)} \tag{3.155}$$

The D.C. current gain of the transistor α_0 is obtained from the above expression by setting $\omega = 0$

$$\alpha_0 = \frac{\tau_n}{\tau_{n1}} \tag{3.156}$$

α cut off frequency

The current gain of the transistor falls to $\frac{1}{\sqrt{2}}$ times its low frequency value at a frequency

$$\omega_{C\alpha} = \frac{1}{\tau_n} \approx \frac{1}{\tau_d} \tag{3.157}$$

We have from equation (3.109) that the current gain in common base configuration has a value

$$h_{21} = -\alpha(1-j\omega\tau_d) \tag{3.109}$$

$$\approx -\alpha/(1+j\omega\tau_d) \tag{3.158}$$

The common base current gain as obtained from the diffusion equations falls to $\frac{1}{\sqrt{2}}$ times its low frequency value at a frequency

$$\omega_{C\alpha} = \frac{1}{\tau_d}$$

The frequency at which the α of a transistor falls to $\frac{1}{\sqrt{2}}$ times its low frequency value is called the α cut off frequency.

β cut off frequency

The frequency dependence of β can also be found as follows:

Writing

$$\beta = \frac{\alpha}{1-\alpha}$$

we have

$$\beta = \frac{\alpha}{1-\alpha} = \frac{\tau_n}{\tau_{n1}(1+j\omega\tau_n)-\tau_n} \tag{3.159}$$

At low frequencies $\beta_0 = \dfrac{\tau_n}{\tau_{n1} - \tau_n}$ (3.160)

and $\beta = \beta_0 \dfrac{1}{1 + j\omega\tau_{nr}}$ (3.161)

The beta cut off frequency or the frequency at which beta reduces to $\frac{1}{\sqrt{2}}$ times its low frequency value is now obtained as

$$\omega_{C\beta} = \frac{1}{\tau_{nr}} \tag{3.162}$$

$$= \frac{1}{\tau_n} - \frac{1}{\tau_{n1}} = \omega_{C\alpha}(1 - \alpha)$$

$$\omega_{C\beta} = \frac{1-\alpha}{\tau_d} \tag{3.162}$$

We have earlier obtained the beta cut off frequency as

$$\omega_{C\beta} = \frac{1}{\beta_0 \tau_d}$$

Equation (3.162) is in agreement with equation (3.132) as

$$\beta_0 = \frac{1}{1-\alpha}$$

3.10.1 *Large signal theory*

Till now we considered the operation of a transistor in A.C. circuits where the alternating voltages are small compared to the steady voltages across the junctions. Transistors are extensively used for switching operations when the voltage variations across the junctions are very large and the simplifying assumptions made earlier are not valid.

The expressions for emitter and collector currents of a transistor used in switching were first derived by Ebers and Moll. The expressions are valid for large signals and are referred to as Ebers-Moll equations. From the expressions for the emitter and collector currents the off and on impedances can be calculated. We derive the expressions on the basis of charge control approximation and alternatively from the concept of four pole parameters.

The charge crossing the emitter base junction depends on the minority carrier density at the edge of the depletion layer which in turn depends on the applied voltage. We have studied under junctions that the hole density on the *N*-side of a forward biased junction is given by the formula $p_1 = p_0 \left(\exp \frac{qV}{kT} - 1 \right)$, where p_0 is the equilibrium value of the hole density and V the applied voltage. As the charge crossing the junction is proportional to the minority carrier density at the edge of the depletion layer we can write

$$Q_E = Q_{ES} \left(\exp \frac{qV_E}{kT} - 1 \right) \tag{3.163}$$

where Q_{ES} is a constant and V_E the voltage across the emitter base junction.

Having obtained the relation between the charge crossing the junction and the voltage across the junction we can write down expressions for the emitter and collector currents in terms of the voltages as

$$\left.\begin{aligned} I_E &= \frac{Q_E}{\tau_n} = \frac{Q_{ES}}{\tau_n} \cdot \left(\exp \frac{qV_E}{kT} - 1 \right) \\ I_C &= \frac{-Q_E}{\tau_{n1}} = \frac{-Q_{ES}}{\tau_{n1}} \left(\exp \frac{qV_E}{kT} - 1 \right) \end{aligned}\right\} \tag{3.164}$$

If the transistor is operated in the reverse manner, collector emitting and emitter collecting we can write expressions for charges and currents as

$$Q_C = Q_{CS} \left(\exp \frac{qV_C}{kT} - 1 \right)$$

$$I_C = Q_C/\tau_i$$

$$I_E = -Q_C/\tau_{i1} \tag{3.165}$$

The reverse current gain

$$\alpha_i = \tau_i/\tau_{i1}$$

τ_i and τ_{i1} are the time constants when the transistor is operated in the reverse mode.

With arbitrary voltages V_E and V_C across the emitter base and collector base junctions the emitter and collector currents of the transistor are

$$\left.\begin{aligned} I_E &= \frac{Q_E}{\tau_n} - \frac{Q_C}{\tau_{i1}} \\ I_C &= \frac{Q_C}{\tau_i} - \frac{Q_E}{\tau_{n1}} \end{aligned}\right\} \tag{3.166}$$

When the emitter is open and the collector base junction is reverse biased, the current flowing in the collector is the reverse saturation current I_{CO}. From expressions (3.164), (3.165) and (3.166) we obtain

$$I_{CO} = \frac{-Q_{CS}}{\tau_i}\,(1-\alpha_n\,\alpha_i) \tag{3.167}$$

Similarly the emitter saturation current with collector open

$$I_{EO} = \frac{-Q_{ES}}{\tau_n}\,(1-\alpha_n\,\alpha_i) \tag{3.168}$$

Substituting the values of Q_{ES} and Q_{CS} from expressions (3.167) and (3.168) in expressions (3.165) and (3.166) we obtain the Ebers-Moll equations for large signals as

$$\left.\begin{aligned} I_E &= \frac{-I_{EO}}{1-\alpha_n\,\alpha_i}\left(\exp\frac{qV_E}{kT}-1\right)+\frac{\alpha_i\,I_{CO}}{1-\alpha_n\alpha_i}\left(\exp\frac{qV_C}{kT}-1\right)\\ I_C &= \frac{\alpha_n\,I_{EO}}{1-\alpha_n\,\alpha_i}\left(\exp\frac{qV_E}{kT}-1\right)-\frac{I_{CO}}{1-\alpha_n\,\alpha_i}\left(\exp\frac{qV_C}{kT}-1\right)\end{aligned}\right\} \tag{3.169}$$

Alternatively we can derive the above expressions from the fact that the emitter and collector currents of a transistor are determined by the junction voltages V_E and V_C and can be expressed as

$$\left.\begin{aligned} I_E &= A_{11}\left(\exp\frac{qV_E}{kT}-1\right)+A_{12}\left(\exp\frac{qV_C}{kT}-1\right)\\ I_C &= A_{21}\left(\exp\frac{qV_E}{kT}-1\right)+A_{22}\left(\exp\frac{qV_C}{kT}-1\right)\end{aligned}\right\} \tag{3.170}$$

A_{11}, A_{12}, A_{21} and A_{22} are a set of constants determining the behaviour of the transistor. If α_n and α_i are the current gains in the normal and reverse operations and I_{CO}, I_{EO} the reverse saturation currents of collector and emitter junctions

$$\left.\begin{aligned} I_E &= -\alpha_i\,I_C-I_{EO}\left(\exp\frac{qV_E}{kT}-1\right)\\ I_C &= -\alpha_n\,I_E-I_{CO}\left(\exp\frac{qV_C}{kT}-1\right)\end{aligned}\right\} \tag{3.171}$$

From equations (3.170) and (3.171) the constants A_{11} etc. can be found as

$$A_{11} = \frac{-I_{EO}}{1-\alpha_n\,\alpha_i}$$

$$A_{12} = \frac{\alpha_i\,I_{CO}}{1-\alpha_n\,\alpha_i}$$

$$A_{21} = \frac{\alpha_n\,I_{EO}}{1-\alpha_n\,\alpha_i}$$

$$A_{22} = \frac{-I_{CO}}{1-\alpha_n\,\alpha_i}$$

Substitution of constants A_{11} etc. in equation (3.170) give us the Ebers-Moll expressions (3.169).

3.10.2 *Transistor as a switch*

The basic circuit of a transistor switch is given in Fig. 3.12. In the off state the collector and emitter junctions are both reverse biased and the current flowing through the transistor is very small. The current can be calculated from expression (3.169) by setting $V_E \leqslant 0$ and $V_C \leqslant 0$.

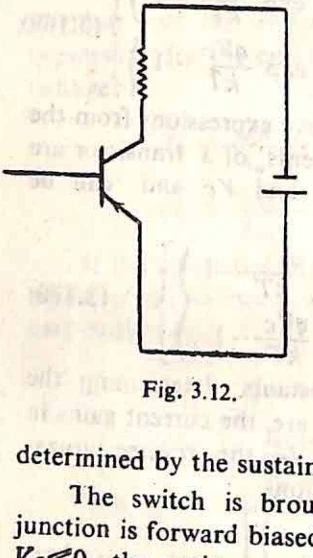


Fig. 3.12.

Under these conditions the collector current

$$I_{C(\text{off})} = \frac{I_{CO} - \alpha_n I_{EO}}{1 - \alpha_n \alpha_i}$$

Applying the auxiliary condition $\alpha_n I_{EO} = \alpha_i I_{CO}$ from symmetrical considerations, we have

$$I_{C(\text{off})} = \frac{I_{CO}(1-\alpha_i)}{1-\alpha_n \alpha_i} \tag{3.172}$$

The off impedance of the switch

$$\frac{V_C}{I_{C(\text{off})}} = \frac{V_C(1-\alpha_n \alpha_i)}{I_{CO}(1-\alpha_i)} \approx \frac{V_C}{I_{CO}} \tag{3.173}$$

The off impedance is very large. The maximum collector potnteial is determined by the sustaining voltage of the transistor.

The switch is brought into active state when the emitter junction is forward biased. In this region of operation $V_E > 0$ and $V_C \leqslant 0$, the emitter and collector currents are obtained from expressions (3.169) as

$$\left.\begin{aligned} I_E &= \frac{-I_{EO}}{1-\alpha_n \alpha_i} \cdot \exp \frac{qV_E}{kT} - \frac{\alpha_i I_{CO}}{1-\alpha_n \alpha_i} \\ I_C &= \frac{\alpha_n I_{EO}}{1-\alpha_n \alpha_i} \cdot \exp \frac{qV_E}{kT} + \frac{I_{CO}}{1-\alpha_n \alpha_i} \end{aligned}\right\} \tag{3.174}$$

The base current when the transistor is in the active state

$$\begin{aligned} I_B &= -(I_E + I_C) \\ &= \frac{I_{EO}(1-\alpha_n)}{1-\alpha_n \alpha_i} \exp \frac{qV_E}{kT} - \frac{I_{CO}(1-\alpha_i)}{1-\alpha_n \alpha_i} \end{aligned} \tag{3.175}$$

Substituting the value of I_B in the expression for I_C we obtain the relation between I_C and I_B as

$$I_C = \frac{I_B \alpha_n}{1-\alpha_n} + \frac{I_{CO}}{1-\alpha_n} \tag{3.176}$$

The collector current of the transistor increases with the base current according to the above expression and the switch is in the active state. As the collector current increases the collector potential rises due to the potential drop in the resistor in series with the collector. The switch reaches the ON state when the collector potential rises to a positive value forward biasing the collector base junction. In this region the transistor is said to be saturated.

The currents and voltages in the saturation region of the transistor can easily be obtained from expressions (3.169) by setting $V_E \gg 0$ and $V_C \gg 0$. The currents and voltages are

$$\left.\begin{aligned}
I_E &= \frac{-I_{EO}}{1-\alpha_n\,\alpha_i}\exp\frac{qV_E}{kT} + \frac{\alpha_i\,I_{CO}}{1-\alpha_n\,\alpha_i}\exp\frac{qV_C}{kT}\\
I_C &= \frac{\alpha_n\,I_{EO}}{1-\alpha_n\,\alpha_i}\exp\frac{qV_E}{kT} - \frac{I_{CO}}{1-\alpha_n\,\alpha_i}\exp\frac{qV_C}{kT}\\
V_E &= \frac{kT}{q}\ln -\frac{I_E+\alpha_i\,I_C}{I_{EO}}\\
V_C &= \frac{kT}{q}\ln -\frac{I_C+\alpha_n\,I_E}{I_{CO}}
\end{aligned}\right\} \tag{3.177}$$

As the transistor is used in the common emitter configuration

$$I_B = -(I_E + I_C)$$

and

$$V_{CE} = (V_C - V_E)$$

The voltage across the transistor in the saturation region

$$V_{CE(\text{sat})} = \frac{kT}{q}\ln\frac{I_{EO}}{I_{CO}}\;\frac{\alpha_n-(1-\alpha_n)\,I_C/I_B}{1+(1-\alpha_i)\,I_C/I_B} \tag{3.178}$$

The saturation voltage depends on the ratio I_C/I_B. This ratio called the D.C. beta is not a constant but varies with the value of I_C. To this saturation voltage the potential drops in the bulk semiconductor are to be added. The voltage across the terminals of the transistor is a fraction of a volt and the ON resistance is about a few ohms.

3.10.3 *Switching times*

When the transistor is in the off state both the emitter and collector junctions are reverse biased. The switch on process involves bringing the transistor into the active state and the time interval between the application of the base current pulse and the transistor entering the active state is called the delay time. It depends on the time constant associated with the input resistance and the transition capacitance of the junctions. The delay time is quite short.

Once the transistor enters the active region the collector current starts increasing, resulting in a fall in collector potential. The turn on time is defined as the time taken for the collector current to increase from 10 to 90% of the saturation value. It was calculated in 'Introduction to Electronics' from the equivalent circuit of the transistor. We now evaluate the turn on time using the charge control approximation.

If Q_E is the charge crossing the emitter junction we have already found

$$\frac{dQ_E}{dt}=I_E-\frac{Q_E}{\tau_n} \tag{3.151}$$

and

$$I_C=\frac{-Q_E}{\tau_{n1}} \tag{3.144}$$

The transform equations for the above are

$$I_E(s)=Q_E(s)\left(s+\frac{1}{\tau_n}\right)$$

$$I_C(s)=\frac{-Q_E(s)}{\tau_{n1}}$$

From the above, after rearrangement, we obtain

$$\frac{I_E(s)+I_C(s)}{I_C(s)}=\frac{-I_B(s)}{I_C(s)}=1-\tau_{n1}\left(s+\frac{1}{\tau_n}\right)$$

and

$$I_C(s)=\frac{I_B(s)}{\tau_{n1}\left(s+\dfrac{1}{\tau_n}\right)-1}$$

As I_B is a step function we can write $I_B(s)=-\frac{I_0}{s}$ and

$$I_C(s)=-\frac{I_0}{s\,\tau_{n1}\left(s+\dfrac{1}{\tau_n}\right)-1}$$

$$=-\frac{I_0}{\tau_{n1}}\cdot\frac{1}{s\left(s+\dfrac{1}{\tau_n}-\dfrac{1}{\tau_{n1}}\right)}$$

As $$\frac{1}{\tau_n} - \frac{1}{\tau_{n1}} = \omega_{C\beta}$$

$$I_C(s) = \frac{I_0}{\tau_{n1}} \cdot \frac{1}{s\,(s+\omega_{C\beta})}$$

The inverse transform of the above equation is

$$I_C = -\frac{I_0\alpha}{1-\alpha} \cdot (1-\exp \omega_{C\beta}\, t) \qquad (3.179)$$

The collector current increases exponentially with time and the time constant involved is $1/\omega_{C\beta}$. The rise time therefore is

$$t_r = 2.2\tau = \frac{2.2}{\omega_{C\beta}} \qquad (3.180)$$

The rise time evaluated above does not take into consideration the Miller capacitance. When the Miller capacitance is taken into account we obtain the expression

$$t_r = 2.2\left(\frac{1}{\omega_{C\beta}} + \beta R_L C_{bc}\right)$$

If the base current is large and the pulse is sufficiently long the transistor is driven into saturation. In the saturation region the collector junction is also forward biased and the stored base charge increases depending on the base current.

The off transition of the transistor switch first involves the removal of the excess base charge bringing the transistor out of saturation. The interval between switching off and the transistor coming out of saturation is called the storage time. The storage time depends on the extent to which the transistor is driven into saturation. After the transistor comes out of saturation the collector current decays exponentially with a time constant same as for the rise. The decay time is the time taken for the collector

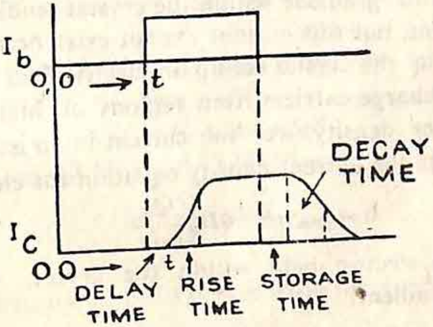


Fig. 3.13.

current to fall from 90 to 10% of the maximum value. The base current and the collector current pulses are shown in Fig. 3.13.

3.11.1 *Limitations of a diffusion transistor*

The transistor we have studied has a base region in which the impurity distribution is uniform and the minority carriers injected into the base from the emitter diffuse across the base to the collector. As the high frequency limit of performance of a transistor is determined by the transit time of minority carriers through the base, a plain diffusion transistor cannot be used at very high frequencies.

The base region resistivity is optimised so that the base spreading resistance may not be large. Under the normal operating conditions the collector depletion layer moves into the base and the punch through voltage in case of a diffusion transistor is small. The collector capacitance and the base width modulation factor $\partial W/\partial V_C$ are large and the transistor cannot be used at high frequencies.

The high frequency limit of a transistor can be raised by a nonuniform impurity distribution in the base. There are various types of transistors using nonuniform impurity distribution in the base. The transistors differ in geometry and are called Mesa, Microalloyed depending on the construction. These transistors are given a common name drift transistors as the minority carriers drift in the base. We will start a study of these transistors with a study of an inhomogeneously doped crystal.

Consider an N-type semiconductor crystal in which the impurity density varies from point to point. As the impurities are all ionised the equilibrium electron density in the crystal is not constant. The concentration gradient within the crystal tends to give rise to a diffusion current, but this current cannot exist because the fixed ionised atoms in the crystal set up an electric field which opposes the diffusion of charge carriers from regions of higher density to regions of lower density. As the current in an isolated crystal is zero we can write the current density equation for electrons as

$$0 = q\mu_n nE + qD_n \frac{dn}{dx} \tag{3.181}$$

where E is the electric field within the crystal and dn/dx the concentration gradient.

From equation (3.181) we find the electric field within the crystal

$$E=-\frac{D_n}{\mu_n}\frac{1}{n}\frac{dn}{dx}=-\frac{kT}{q}\frac{1}{n}\frac{dn}{dx} \tag{3.182}$$

The electric field within the crystal will have a constant value when the right hand side of equation (3.182) is constant

$$\frac{kT}{q}\frac{1}{n}\frac{dn}{dx}=-\text{constant};\quad \frac{dn}{n}=-\text{const}\cdot dx$$

Integrating the expression above we find that

$$n=a\exp(-bx) \tag{3.183}$$

where a and b are constants. From equation (3 183) it is obvious that the electric field within the crystal will have a constant value when the impurity distribution within the crystal varies exponentially with distance. As we have set the electric field E to be a constant we immediately find

$$\frac{dE}{dx}=\frac{d^2V}{dx^2}=0 \tag{3.184}$$

From the Poisson's equation we infer that the crystal is space charge neutral point by point.

3.11.2 *Conditions in the base of a transistor when the impurity distribution changes exponentially*

We studied earlier that in the active state of a *PNP* transistor the hole density has a maximum value at the emitter side of the base and a zero value at the collector side and the holes diffuse through the base. When the impurity distribution in the base is not uniform there is a built-in field and the holes drift under the influence of the field. If we assume the density of donors on the emitter side of the base to have a value N_1 the donor density at a distance x from the emitter side can be written as

$$N_x=N_1\exp(-bx) \tag{3.185}$$

where b is a constant. If N_2 is the donor density on the collector side of the base and W the base width

$$N_2=N_1\exp(-bW) \tag{3.186}$$

From equations (3.185) and (3 186) we get

$$\ln\frac{N_1}{N_2}=bW$$

writing $bW=\eta$.

The donor density at a distance x from the emitter edge of the base

$$N_x = N_1 \exp(-\eta/W)x \tag{3.187}$$

and the field within the base region

$$E = \frac{kT}{q} \quad \frac{\eta}{W} \tag{3.188}$$

η is thus measure of the constant built-in field within the base.

In a diffusion transistor holes diffuse from the emitter side to the collector side and the diffusion current density of holes is given by the expression

$$J_p = -qD_p \frac{dp}{dx}$$

When there is a built-in field within the base the holes drift under the influence of the field and the hole current in the base is the sum of the drift. and diffusion currents. At low injection levels the change in electron density within the base is very small compared to its equilibrium value and the built-in electric field within the base region may be considered to remain constant at the value given by the expression (3.188). However at high injection levels the electron density changes appreciably resulting in a change in the magnitude of the built-in field. If the recombination in the base region is neglected and the hole current density is assumed to remain constant within the base, we can write an expression for the hole current in the base region as

$$J_p = -qD_p \frac{dp}{dx} + q\mu_p pE \tag{3.189}$$

Substituting the value of E from equation (3.188) into (3.189) we get, after rearranging

$$J_p = -qD_p \left(\frac{dp}{dx} - p\frac{\eta}{W} \right) \tag{3.190}$$

Solving the above differential equation with the limits $p=0$ at $x=W$

$$p = \frac{J_p W}{qD_p} \cdot \frac{1-\exp\left[-\eta\left(1-\frac{x}{W}\right)\right]}{\eta} \tag{3.191}$$

The hole density varies with the distance according to the above expression. On differentiating the above expression and rearranging.

We have

$$-qD_p \frac{dp}{dx} = J_p \exp\left[-\eta\left(1-\frac{x}{W}\right)\right] \tag{3.192}$$

The left hand side of the above equation is the diffusion current through the base and it is a fraction of the total current density The fraction can be evaluated at different values of x/W and for different values of η

$\frac{x}{W}$	$\exp\left[-\eta\left(1-\frac{x}{W}\right)\right]$		
	$\eta = 0$	$\eta = 1$	$\eta = 4$
0.0	1	0.36	0.02
0.2	1	0.45	0.04
0.4	1	0.55	0.09
0.6	1	0.67	0.2
0.8	1	0.82	0.45
1.0	1	1.0	1.0

From the above table we see that whereas the current is entirely due to diffusion in a transistor with uniform base doping ($\eta = 0$) the fraction of current due to diffusion decreases as the impurity gradient increases. The minority carriers drift under the influence of the built-in field and the transistors with inhomogeneous base are called drift transistors.

The high frequency limit of a transistor is determined by the transit time of minority carriers through the base region of the transistor. A quantitative idea about the improvement in high frequency response is obtained by evaluating the transit time of carriers through the base region of transistor. The stored base charge in the base region is obtained from equation (3.191) as

$$Q_D = \int_0^W qpA\, dx = \frac{J_p WA}{\eta D_p} \int_0^W 1-\exp\left[-\eta\left(1-\frac{x}{W}\right)\right] dx$$

$$= \frac{I_p W^2}{D_p}\left[\frac{\eta-1+\exp(-\eta)}{\eta^2}\right] \tag{3.193}$$

The transit time of carriers through the base for a drift transistor

$$\tau_d = \frac{Q_D}{I_p} = \frac{W^2}{D_p}\left[\frac{\eta-1+\exp(-\eta)}{\eta^2}\right] \tag{3.194}$$

The transit time of holes through the base region of a diffusion transistor is given by the expression

$$\tau_d = \frac{W^2}{2D_p} \tag{3.98}$$

It is easily seen that expression (3.194) reduces to (3.98) as η

tends to zero. The transit time for the same base width is reduced by a factor $2\left[\frac{\eta-1+\exp(-\eta)}{\eta^2}\right]$ in a drift transistor. However, η cannot be increased beyond a value 6 to 8 as increase in η makes the electron density on the base side of the emitter junction very large reducing the emitter efficiency. The donor density in a drift transistor decreases from a large value on the emitter side of the base to almost intrinsic value at the collector side of the base.

3.11.3 *Collector base depletion layer*

Let the donor density at the collector side of the base be N_2. The density increases exponentially as we move towards the emitter side of the base and the donor density at a distance x from the edge of the base on the collector side can be written as

$$N = N_2 \exp - \frac{\eta}{W} \cdot x \qquad (3.194)$$

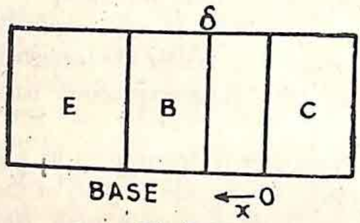


Fig. 3.14.

When a negative voltage is applied to the collector the depletion layer spreads into the base as the collector region is heavily doped. Applying Poisson's relation we have. within the depletion layer

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\epsilon} = \frac{-qN}{\epsilon} = \frac{-qN_2 \exp\left(-\frac{\eta}{W}x\right)}{\epsilon} \qquad (3.195)$$

Integrating we have

$$\frac{dV}{dx} = \frac{qN_2}{\epsilon} \cdot \frac{W}{\eta} \exp\left(-\frac{\eta}{W}x\right) + \text{constant} \qquad (3.196)$$

If the depletion layer extends by an amount δ into the base we have

$$\frac{dV}{dx} = -E \quad \text{when } x = -\delta$$

The constant in equation (3.196) has a value

$$\text{Constant} = -E - \frac{qN_2}{\epsilon} \cdot \frac{W}{\eta} \cdot \exp \frac{\eta}{W} \cdot \delta$$

$$\frac{dV}{dx} = \frac{W}{\eta} \cdot \frac{qN_2}{\epsilon}\left(\exp\frac{-\eta x}{W} - \exp\frac{\eta\delta}{W}\right) - E \qquad (3.197)$$

The potential difference existing across the depletion layer can be found by integrating the expression (3.197) between the limits $x = 0$ and $x = -\delta$

$$V = \frac{W^2}{\eta^2}\,\frac{qN_2}{\epsilon}\left[1 + \frac{\eta\delta}{W}\exp\frac{\eta\delta}{W} - \exp\frac{\eta\delta}{W}\right] + E\delta \qquad (3.198)$$

If we assume that the depletion layer extends by a small distance into the base so as to make $\frac{\eta\delta}{W}$ small the exponential can be expanded to give

$$V = \frac{qN_2\delta^2}{\epsilon} + E\delta \qquad (3.199)$$

Solving the quadratic in δ we have

$$\delta = -\frac{\epsilon E}{qN_2} - \sqrt{\frac{\epsilon^2E^2}{q^2N_2{}^2} + \frac{2V\epsilon}{qN_2}} \qquad (3.200)$$

The width of the depletion layer for $\eta = 0$ (uniformly doped case) is

$$\delta = -\sqrt{\frac{2V\epsilon}{qN_2}} \qquad (3.201)$$

We thus see that the depletion layer is wider in case of a drift transistor. When the width of the depletion layer increases the collector capacitance decreases as the capacitance is given by the expression $C = \frac{\epsilon A}{d}$. The collector capacitance for the drift transistor is less than for the diffusion transistor.

3.11.4 *Base width modulation factor*

We have seen earlier in this chapter that when a transistor is used in A.C. circuits the changes in collector voltage cause changes in base width and the base width change with collector voltage, appears in the A.C. impedances of the transistor in the form $\frac{\partial W}{\partial V_C}$. We will now evaluate this factor for a drift transistor. The electrical base width W of a transistor is less than the geometrical base width W_0 and

$$W = W_0 - \delta \qquad (3.202)$$

Differentiating with respect to V (the collector base voltage) we find

$$\frac{\partial W}{\partial V_C} = -\frac{\partial \delta}{\partial V_C} \qquad (3.203)$$

Differentiating equation (3.200) we have

$$\frac{\partial \delta}{\partial V_C} = -\left(\frac{\epsilon^2 E^2}{q^2 N_2^2} + \frac{2V\epsilon}{qN_2}\right)^{-1/2} \cdot \frac{\epsilon}{qN_2}$$

$$\frac{\partial W}{\partial V_C} = \frac{\epsilon}{qN_2\left(\frac{\epsilon^2 E^2}{q^2 N_2^2} + \frac{2V\epsilon}{qN_2}\right)^{1/2}}$$

$$= \frac{1}{\left(E^2 + \frac{2VqN_2}{\epsilon}\right)^{1/2}} \tag{3.204}$$

For a diffusion transistor $E = 0$

and
$$\frac{dW}{dV_C} = \left(\frac{\epsilon}{2VqN_2}\right)^{1/2} \tag{3.205}$$

Thus we see that the factor $\frac{dW}{dV_C}$ is smaller for the drift transistor.

3.11.5 *Punch through voltage*

In this section we will calculate the value of the punch through voltage for a drift transistor. When the punch through occurs the depletion layer extends throughout the base region, this voltage can be easily found by integrating the expression (3.197) between the limits 0 and $-W$. Integrating we find

$$V = \frac{N_2 W^2 q}{\epsilon} \frac{\eta - 1 + \exp(-\eta)}{\eta^2} + EW \tag{3.206}$$

EW is the potential difference required to overcome the built-in field and this is very small and can be neglected in comparison with the first term. The punch through voltage for a drift transistor is thus

$$V = \frac{N_2 W^2 q}{\epsilon} \frac{\{\eta - 1 + \exp(-\eta)\}}{\eta^2} \tag{3.207}$$

The punch through voltage for a diffusion transistor can be found by setting $\eta = 0$; when this is done it is seen that the punch through voltage has a value

$$V = \frac{N_2 W^2 q}{2\epsilon} \tag{3.208}$$

Comparing equations (3.207) and (3.208) it is obvious that the punch through voltage for a drift transistor is higher than for a diffusion transistor.

Summarising, the drift transistor has the following advantages over a diffusion transistor:

1. Lesser transit time.
2. Lesser collector capacitance.
3. Small value for dW/dV_C.
4. Larger punch through voltage.

However the drift transistor has the following disadvantage:

1. The emitter efficiency is lower as the doping on the emitter side of the base is large.

Important Expressions

$$\text{Emitter Efficiency} = \frac{1}{1+\frac{\sigma_B}{\sigma_E}\,\frac{L_p}{L_n}\tanh\frac{W}{L_p}}$$

$$= \frac{1}{1+\frac{\sigma_B}{\sigma_E}\cdot\frac{W}{L_n}}$$

$$\text{Transport Factor} = \operatorname{sech}\frac{W}{L_p}$$

$$= 1-\left(\frac{W}{L_p}\right)^2$$

$$\text{Collector Multiplication Factor} = \frac{1}{1-\left(\frac{V}{V_{br}}\right)^n}$$

$$I_{CEO} = \frac{I_{CO}}{1-\alpha}$$

$$\text{Sustaining Voltage} = V_{br}(1-\alpha)^{1/n}$$

$$\text{Stored Base Charge, } Q_s = \frac{W^2}{2D_p}\cdot I_E$$

y-parameters in Common Base Configuration

$$y_{11} = \frac{qI_E}{kT}\left(1+j\omega\alpha\frac{W^2}{3D_p}\right)$$

$$y_{21} = \frac{\alpha qI_E}{kT}\left(-1+j\omega\frac{W^2}{6D_p}\right)$$

$$y_{22} = \frac{\alpha I_E}{W}\cdot\frac{\partial W}{\partial V_C}\left(1+j\omega\frac{W^2}{3D_p}\right)$$

$$y_{12} = \frac{\alpha I_E}{W}\,\frac{\partial W}{\partial V_C}\left(-1+j\omega\frac{W}{6D_p}\right)$$

Transit time of holes through base region $= \tau_d = \dfrac{W^2}{2D_p}$.

Common Emitter h-parameters

$$h_{11} = \frac{1}{g_1(1-\alpha)+j\omega g_1\tau_d}$$

$$h_{12} = \frac{g_2}{g_1(1-\alpha)} \cdot j\omega\tau_d$$

$$h_{21} = \frac{\alpha}{1-\alpha}\,(1-j\omega\tau_d)$$

$$h_{22} = g_2\left(1+\frac{\alpha}{1-\alpha}\,j\omega\tau_d\right)$$

$$g_m = g_1\alpha+j\beta g_1\tau_d$$

α cut off frequency $\omega_{C\alpha} = \dfrac{1}{\tau_d}$

β cut off frequency $\omega_{C\beta} = \dfrac{1}{\beta_0\tau_d}$

Ebers-Moll Equations

$$I_E = \frac{-I_{EO}}{1-\alpha_n\alpha_i}\left(\exp\frac{qV_E}{kT}-1\right)+\frac{\alpha_i I_{CO}}{1-\alpha_n\alpha_i}\left(\exp\frac{qV_C}{kT}-1\right)$$

$$I_C = \frac{\alpha_n I_{EO}}{1-\alpha_n\alpha_i}\left(\exp\frac{qV_E}{kT}-1\right)-\frac{I_{CO}}{1-\alpha_n\alpha_i}\left(\exp\frac{qV_C}{kT}-1\right)$$

OFF Impedance of a Transistor Switch $= \dfrac{V_C(1-\alpha_n\alpha_i)}{I_{CO}(1-\alpha_i)}$

$$V_{CE(\text{sat})} = \frac{kT}{q}\ln\frac{I_{EO}}{I_{CO}}\,\frac{\alpha_n-(1-\alpha_n)I_C/I_B}{1+(1-\alpha_i)I_C/I_B}$$

Rise Time $t_r = 2.2\tau = \dfrac{2.2}{\omega_{C\beta}}$

Transit Time of Minority Carriers in a Drift Transistor

$$\tau_d = \frac{W^2}{D_p}\left\{\frac{\eta-1+\exp(-\eta)}{\eta^2}\right\}$$

Questions

1. Derive an expression for the emitter efficiency of a transistor. How can the emitter efficiency be increased?
2. What is meant by transport factor? Derive an expression for the transport factor and explain how it can be increased.

3. Draw a neat circuit diagram of the set up you use for determining the common base characteristics of a transistor. Draw typical curves and give necessary interpretation.
4. Why is the current in collector larger than the reverse saturation current when the base lead is open?
5. Discuss the characteristics of a transistor in the common emitter configuration.
6. What is (a) punch through, (b) sustaining voltage.
7. Show that two-thirds of the stored charge resides on the junction capacitance of a transistor.
8. Assuming that two-thirds of stored charge resides on the junction capacitance evaluate the common base y parameters y_{11} and y_{21}.
9. Explain how Early effect gives rise to the output admittance. Evaluate the common base y parameters y_{22} and y_{12}.
10. Evaluate the transit time of minority carriers through the base region of a transistor.
11. From the defining equations of the y parameters construct an equivalent circuit of a transistor.
12. Define the various 'h' parameters and relate them to the various elements in the equivalent circuit of a transistor in the common base configuration.
13. From the defining equations obtain the relationships between common base and common emitter y parameters.
14. Draw the equivalent circuit of a transistor in the common emitter configuration and relate the various components of the equivalent circuit to the 'h' parameters.
15. How does the common emitter current gain vary with frequency? Obtain the relation between the cut off frequency and the transit time.
16. Show that the transit time of minority carriers reduces by 50% when the current through the transistor increases to a high value. Does it mean that the current gain increases considerably when a transistor is operated at high currents?
17. What do you understand by thermal resistance of a transistor?
18. Evaluate the α and β cut off frequencies using the charge control concept.
19. Derive Ebers-Moll equations for a transistor.
20. Evaluate the OFF and ON impedances of a transistor switch.

21. Explain the various switching times of a transistor switch. Evaluate the turn on time.
22. Evaluate the transit time of minority carriers through the base of a drift transistor.
23. Evaluate the base width modulation factor of a graded base transistor and compare it with the base width modulation factor of diffusion transistor.
24. Evaluate the punch through voltage of a graded base transistor and compare it with the punch through voltage of a diffusion transistor.

4. Junction Field Effect Transistor

The transistor studied in the last chapter is based on the control of minority carrier flow within a small region called the base. In a field effect transistor the majority carrier flow in the bulk semiconductor is controlled by varying the sectional area through which carriers flow. This variation in sectional area is accomplished by controlling the movement of the depletion layer into the bulk semiconductor under the influence of a reverse voltage existing between the bulk semiconductor and a neighbouring layer of oppositely doped semiconductor material. The field effect transistor resembles a triode in which the electron flow from the cathode to plate is controlled by the potential on the grid. However, the electrical characteristics of the triode and the field effect transistor do not show much resemblance as the flow of carriers in the transistor is controlled by a change in resistance due to a change in sectional area through which the carriers flow. In the following few pages we will study the construction and some basic properties of the field effect transistor.

4.1 Structure of a Field Effect Transistor

The field effect transistor is manufactured by depositing a thin *N*-type epitaxial layer on a heavily doped *P* substrate. A heavily doped *P* layer is again deposited on the *N* epitaxial layer. Contacts are made to the ends of the *N* filament sandwiched between two heavily doped *P* regions and also to the *P* regions on either side of

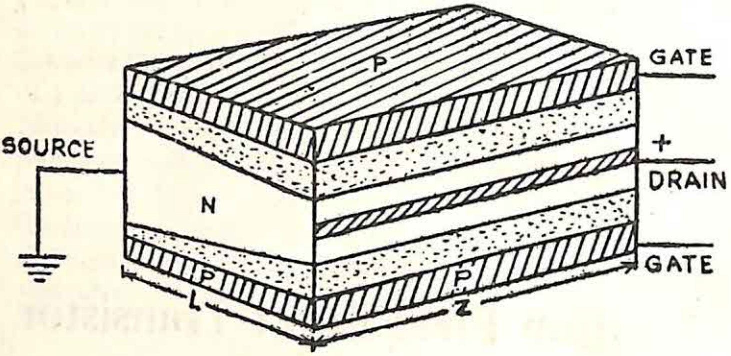


Fig. 4.1.

the *N* filament. The *N* filament is called the channel. Majority carriers—electrons flow from one end of the channel to the other end. The end from where the electrons move is called the cathode or the source and the end of the channel at which the electrons are collected is called the anode or the drain. The *P* regions on either side of the channel are called gates. We now have an *N* channel field effect transistor. By depositing a *P* epitaxial layer on a heavily doped *N* substrate a *P* channel field effect transistor is formed.

4.2 Characteristics of a FET.

The channel current or the current through the transistor depends on the drain and gate potentials which are measured relative to the source. As the semiconductor regions on either side of the channel are heavily doped the depletion layer formed between the *P* and *N* regions spreads into the channel and the width of the depletion layer at any point depends on the potential difference between the layers at the point. When a potential difference is applied to the ends of the channel a current is established through the channel, and the current tends to increase with the drain voltage as in the case of any resistor, but the increase in drain voltage widens the depletion layer and the channel area is reduced. Therefore as the drain voltage increases the channel current also increases but not in direct proportion. This region of operation of the transistor is referred to as ohmic or linear region. When the drain voltage reaches a value so that the depletion layer spreads throughout the channel, the channel is pinched off and the current saturates. After the channel is pinched off the channel current is almost independent of the gate voltage till the transistor breaks down. The drain potential controls the pinch off as it directly affects the voltage across the *P* and *N* regions and hence controls the width of the depletion layer. The saturation current is thus a function of the gate potential. The variation of the saturation current with the gate potential enables the use of FET as a control device. The device is also operated in the ohmic region when it functions as a controlled resistor.

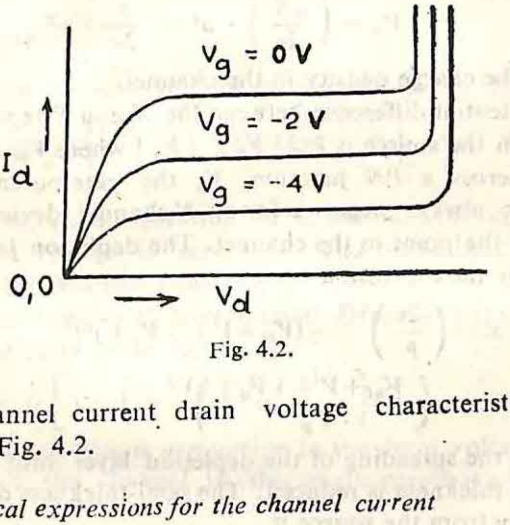


Fig. 4.2.

The channel current drain voltage characteristics are as indicated in Fig. 4.2.

4.3 *Analytical expressions for the channel current*

The variation of the channel current with the drain and gate potentials can easily be calculated for a symmetrical transistor in

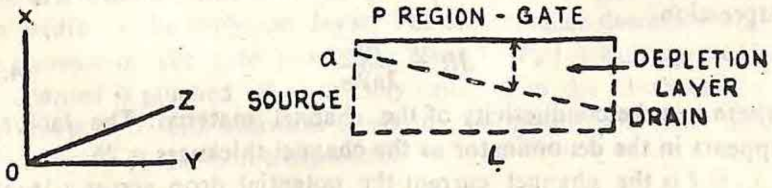


Fig. 4.3.

which the gate regions are supposed to extend over equal areas on either side of the channel. As the areas are symmetrical we can consider half of the channel with a gate on one side.

Let the semi-thickness of the channel be a, length L and width Z.

Remembering that the width of the depletion layer of a P^+N junction is given by expression $x=\sqrt{\frac{2\epsilon}{qN}V}$. We find that the channel pinches off for a voltage V_p such that

$$a=\sqrt{\frac{2\epsilon}{qN}V_p} \tag{4.1}$$

The pinch off voltage is given by the expression

$$V_p=\left(\frac{qN}{2\epsilon}\right)\cdot a^2=\frac{\rho}{2\epsilon}\cdot a^2 \tag{4.2}$$

where ρ is the charge density in the channel.

The potential difference between the N and P regions at a distance y from the source is $V_{bi}+V_y+|V_g|$ where V_{bi} is the built-in potential across a PN junction, V_g the gate potential (the gate potential is always negative for an N channel device) and V_y the potential at the point in the channel. The depletion layer thickness x is given by the expression

$$\begin{aligned} x&=\left(\frac{2\epsilon}{\rho}\right)^{1/2}\cdot(V_{bi}+V+|V_g|)^{1/2} \\ &=a\left(\frac{V_{bi}+V+|V_g|}{V_p}\right)^{1/2} \end{aligned} \tag{4.3}$$

Due to the spreading of the depletion layer into the channel the channel thickness is reduced. The semi-thickness of the channel at distance y from the source is

$$h=(a-x)=a\left[1-\left(\frac{V_{bi}+V_y+|V_g|}{V_p}\right)^{1/2}\right] \tag{4.4}$$

and the resistance of a length dy, the channel dR is given by the expression

$$dR=\frac{dy}{2hZ\sigma} \tag{4.5}$$

where σ is the conductivity of the channel material. The factor 2 appears in the denominator as the channel thickness is $2h$.

If I is the channel current the potential drop across a length dy of the channel dV is obtained as

$$dV = \frac{Idy}{2hZ\sigma}$$

Substituting the value of h from expression (4.4) and rearranging we have

$$Idy = 2Z\sigma a\left[1 - \left(\frac{|V_g| + V_y + V_{bi}}{V_p}\right)^{1/2}\right]dV$$

Integrating along the length of the channel we have

$$IL = 2Z\sigma a\left[V_d - \frac{2}{3V_p^{1/2}}\left\{\left(|V_g| + V_d + V_{bi}\right)^{3/2} - \left(|V_g| + V_{bi}\right)^{3/2}\right\}\right] \tag{4.6}$$

$\frac{2Z\sigma a}{L}$ is the conductance of the N channel when the channel thickness is not changed by the movement of the depletion layer and we can write $\frac{2Z\sigma a}{L} = G$.

Substituting the value of G in expression (4.6) we have

$$I = G\left[V_d - \frac{2}{3V_p^{1/2}}\left\{\left(|V_g| + V_d + V_{bi}\right]^{3/2} - \left[|V_g| + V_{bi}\right]^{3/2}\right\}\right] \tag{4.7}$$

The two regions of operation of interest are the ohmic region when the drain voltage is very small and the saturation region when the current becomes constant.

In the ohmic region V_d is very small. Expanding the bracketed term in expression (4.7) we have

$$I = G\left[1 - \left(\frac{|V_g| + V_{bi}}{V_p}\right)^{1/2}\right]V_d \tag{4.8}$$

The channel current is proportion to the drain voltage and the conductance of the channel in the ohmic region is given by the expression

$$g = G\left[1 - \left(\frac{|V_g| + V_{bi}}{V_p}\right)^{1/2}\right] \tag{4.9}$$

From the above expression it is obvious that the channel conductance is a function of the gate potential which determines the width of the depletion layer. The conductance decreases with the increase of the gate potential. When ($|V_g| + V_{bi}$) equals V_p the channel is pinched off completely resulting in the conductance becoming zero. The transistor is said to be turned off. The turn off voltage is given by the expression

$$V_T = V_p - V_{bi} \tag{4.10}$$

As is evident from the above discussion the FET when operated in the ohmic region behaves as a resistor and the resistance value can be changed by the gate potential.

The current reaches a maximum value when the channel is pinched off and the pinch off occurs when $|V_g| + V_d + V_{bi} = V_p$. Substituting the pinch off condition in (4.7) we obtain the expression for the saturation current

$$I_{\text{sat}} = G\left[V_d - \frac{2}{3}V_p + \frac{2}{3}V_p\left(\frac{V_g + V_{bi}}{V_p}\right)^{3/2}\right] \tag{4.11}$$

4.4 Transconductance

The current in the channel is dependent on the gate potential as the gate potential changes the width of the channel. The transconductance of the transistor is defined as

$$g_m = \frac{\partial I}{\partial V_g}\bigg|_{V_{d(\text{const})}} \tag{4.12}$$

The transconductance of the transistor can be found by differentiating (4.7) as

$$g_m = -G\left[\left(\frac{|V_g| + V_d + V_{bi}}{V_p}\right)^{1/2} - \left(\frac{|V_g| + V_{bi}}{V_p}\right)^{1/2}\right] \tag{4.13}$$

The negative sign for the transconductance indicates that the current decreases with increase in $|V_g|$. This is expected as an increase in the gate potential widens the depletion layer and the channel width is reduced. The transconductance in the ohmic region can be obtained from expression (4.13) by expanding the bracketed term as

$$g_m \approx \tfrac{1}{2}G\frac{V_d}{V_p^{1/2}\left(|V_g| + V_{bi}\right)^{1/2}} \quad \text{(linear region)} \tag{4.14}$$

The transconductance in the saturation region is obtained by setting $(V_g + V_d + V_{bi}) = V_p$ as

$$g_m = -G\left[1 - \left(\frac{|V_g| + V_{bi}}{V_p}\right)^{1/2}\right] \tag{4.15}$$

Comparing expressions (4.15) and (4.9) we notice that the transconductance in the saturation region equals conductance in the linear region.

4.5 Non-uniform Impurity Distribution in the Channel

We have till now considered the voltage current relationship for a field effect transistor with uniform doping density in the channel. Field effect transistors are also manufactured by the process of diffusion just as bipolar transistors, in which case the impurity density in the channel does not remain constant. We now consider an extreme case in which the impurities are concentrated on the side of the channel opposite to the gate layer (delta distribution).

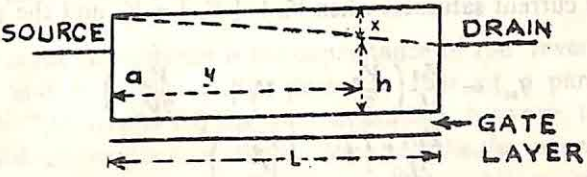


Fig. 4.4.

Consider a channel of width *a* having impurities concentrated on one side of the channel. If ρ is the charge density of the mobile carriers per unit volume in the channel the charge density of ionised impurities on one side of the channel is ρa per unit area. When a depletion layer is formed the field intensity within the depletion layer is $\frac{\rho a}{\epsilon}$ as it is assumed that the charges are concentrated on one side of the channel. The potential V_y at a point distant y from the source is given by the expression

$$V_y = \frac{\rho a}{\epsilon} \cdot x \tag{4.16}$$

The channel pinches off when $x=a$, therefore the pinch off voltage for delta distribution

$$V_p = \frac{\rho a^2}{\epsilon} \tag{4.17}$$

[Compare with the pinch off voltage for uniform distribution expression (4.2).]

The channel width at a distance y from the source is

$$h = a - x = a\left(1 - \frac{V_y}{V_p}\right)$$

and the resistance dR for a length dy of the channel is

$$dR = \frac{\rho dy}{hZ}$$

and the potential change across a length dy when a current I flows through the channel is given by the expression

$$IdR = \frac{I\rho dy}{Za\left(1-\frac{V_y}{V_p}\right)} = dV$$

Rearranging and integrating over the length of the channel we have

$$IL = \frac{Za}{\rho}\left[V_d - \frac{1}{2V_p}\left\{\left(V_d + |V_g|\right)^2 - V_g^2\right\}\right]$$

The current saturates when $V_d + |V_g| = V_p$ and the saturation current

$$I_{\text{sat}} = \frac{Za}{\rho L}\left(\frac{V_p}{2} - |V_g| + \frac{V_g^2}{2V_p}\right)$$

$$= \frac{ZaV_p}{2L\rho}\left(1 - \frac{|V_g|}{V_p}\right)^2 \qquad (4.18)$$

The saturation current is directly proportional to $\left(1-\frac{|V_g|}{V_p}\right)^2$ and the device functions as a square law device.

It has been stated in 'Introduction to Electronics' that the saturation current I_{dsat} is related to the saturation current when the gate potential is zero (I_{dss}) by an expression (2.78)

$$I_{\text{dsat}} = I_{\text{dss}}\left(1-\frac{|V_g|}{V_p}\right)^2$$

The expression immediately follows from expression (4.18).

4.6.1 *Junction breakdown*

When the drain potential is increased to a large value the junction breaks down and the current steeply increases. As the breakdown occurs near the drain where the potential difference across the junction is maximum, the breakdown voltage is related to the gate and drain voltages by the expression

$$V_{br} = |V_g| + V_d \qquad (4.19)$$

4.6.2 *Input impedance*

As the gate is the input terminal the input impedance is the parallel combination of the resistance of a reverse biased junction and the transition capacitance of the junction.

The input current to the device is given by the expression

$$I_g = I_s\left[\exp\frac{qV_g}{nkT} - 1\right] \qquad (4.20)$$

where I_s is the reverse saturation current of the reverse biased junction, n has a value 1 where there is no carrier generation within the depletion layer (n becomes 2 when there is carrier generation). The incremental resistance

$$r_{in}=\left(\frac{\partial V_g}{\partial I_g}\right)=\frac{nkT}{q(I_g+I_s)} \tag{4.21}$$

The incremental resistance is of the order of a few hundreds of megohms.

The input capacitance is the capacitance of the reverse biased junction and its value is the same as that of a parallel plate capacitor. The area is ZL and the separation between the plates is the width of the depletion layer. We can take the average width of the depletion layer $\overline{W}$ and write an expression for the capacitance as

$$C_g=\frac{\epsilon ZL}{\overline{W}} \tag{4.22}$$

The capacitance decreases with the increase in gate and drain potentials. The capacitance is a minimum when the depletion layer spreads into half the channel. The minimum capacitance is therefore

$$C_{g(\min)}=\frac{\epsilon ZL}{a} \tag{4.23}$$

4.6.3 *Response time of a field effect transistor*

When the gate potential changes the charges on the gate side and channel side of the depletion layer change. The change in the charge on the channel side of the depletion layer is associated with a momentary current. The response time is the time during which the momentary current flows. The response time can be calculated as the ratio of the charge to the current. The change in the charge due to a change δV_g in the gate potential is $\delta V_g \cdot C_g$. The channel current due to a change in gate potential is $g_m\,\delta V_g$. Therefore the response time

$$t_0=\frac{C_g\cdot\delta V_g}{g_m\delta V_g}=\frac{C_g}{g_m} \tag{4.24}$$

The high frequency limit of operation of the device is reciprocal of the minimum response time and is therefore obtained as

$$\omega_{\max}=\frac{g_m}{C_{\min}} \tag{4.25}$$

$$C_{\min} = \frac{\epsilon Z L}{a}$$

The transconductance is given by the expression (4.15), its maximum value is the conductance of the channel G

$$g_{m(\max)} = G = \frac{2Z\sigma a}{L} = \frac{2ZN_d \cdot \mu_n a}{L}$$

Substituting the maximum value of g_m and minimum value of C_g in expression (4.25) we obtain the high frequency limit

$$\omega_{\max} = \frac{2N_d \cdot \mu_n a^2}{\epsilon L^2} \tag{4.26}$$

The high frequency limit is determined from the channel dimensions and the mobility. As the mobility of electrons is more than that of holes the high frequency limit of N channel FET's is higher than that of P channel FET's.

4.6.4 *Incremental conductance of the channel*

It was stated earlier that the current reaches a saturation value when the channel is pinched off. When the drain potential changes the pinch off point moves in the channel thereby changing the distance between the source and the pinch off point. The change in the distance between the source and the pinch off point causes a small variation in the drain current. Thus we see that the output conductance has a small but definite value.

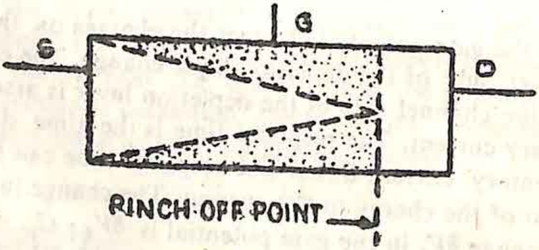


Fig. 4.5.

4.6.5 *Series resistances*

In a practical version of the JFET the channel sandwiched between the depletion layers is away from the contacts as shown in Fig. 4.6. The source to channel resistance R_s and the drain to channel resistance R_d appear in series with the channel.

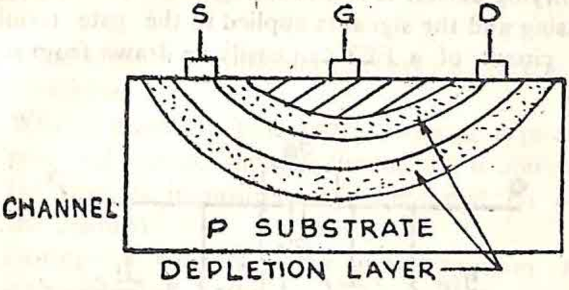


Fig. 4.6.

In the linear region of operation the series resistances add to the resistance of the channel and the observed channel conductance $g_{(obs)}$ in the linear region is related to the channel conductance g by the expression

$$g_{(obs)} = \frac{g}{1+g(R_s+R_d)} \tag{4.27}$$

In the saturation region of operation the drain resistance has practically no effect but the source resistance R_s affects the mutual conductance g_m. This is so because the actual voltage appearing across the gate and channel V_g is the sum of the applied voltage $V_{g(appl)}$ and the potential drop in that part of the channel between the source contact and the edge of the depletion layers on the source side. The actual gate voltage is $V_g = V_{g(appl)}+V_s$, where V_s is the potential drop across the source resistance.

$$V_g = V_{g(appl)} - V_s$$

$$\begin{aligned} g_{m(obs)} &= \frac{\partial I}{\partial V_{g(appl)}} = \frac{dI}{d(V_g+V_s)} \\ &= \frac{1}{\dfrac{dV_g}{dI}+\dfrac{dV_s}{dI}} \\ &= \frac{1}{\dfrac{1}{g_m}+R_s} \\ &= \frac{g_m}{1+g_m R_s} \end{aligned} \tag{4.28}$$

Equivalent circuit of FET

A FET is generally used in the common source configuration as an amplifying device. It is first brought into the active state by proper biasing and the signal is applied to the gate terminal. The equivalent circuit of a FET can easily be drawn from its physical behaviour.

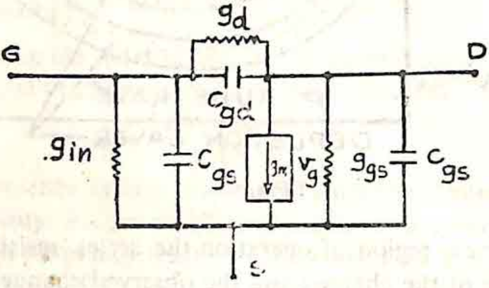


Fig. 4.7.

The g_{in} and C_{gs} represent the input impedance and the reverse biased gate channel junction. The g_{gd} and C_{gd} are the incremental gate drain conductance and the gate drain capacitance. The $g_m V_g$ is the current generator representing changes in channel current caused by changes in gate potential. The g_{ds} and C_{ds} represent the output conductance and capacitance between the drain and source. The output conductance is due to the movement of the pinch off point.

Important Formulae

Pinch off voltage $\frac{\rho}{2\epsilon}a^2$ uniformly doped channel

Delta distribution of impurities in channel $\rho a^2/\epsilon$

Channel Conductance $g = G\left[1-\left(\frac{|V_g|+V_{bi}}{V_p}\right)^{1/2}\right]$

Saturation Current

$$I_{d(\text{sat})} = G\left[V_d-\frac{2}{3}V_p+\frac{2}{3}V_p\left(\frac{V_d+V_{bi}}{V_p}\right)^{3/2}\right]$$

Transconductance $g_m = \frac{1}{2}G\frac{V_d}{V_p^{1/2}[\,|V_g|+V_{bi}]^{1/2}}$

$$g_{m(\text{sat})} = -G\left[1-\left(\frac{V_g+V_{bi}}{V_p}\right)^{1/2}\right]$$

$$r_{in} = \frac{nkT}{q(I_g+I_s)}$$

Maximum Frequency of Operation $2N_a\mu_n a^2/\epsilon L^2$

Questions

1. What is the basic difference between a JFET and a bipolar transistor?
2. What is meant by 'pinch off'? Derive an expression for the pinch off voltage when (a) the channel is uniformly doped, (b) when the impurities are concentrated on one side of the channel.
3. Distinguish between linear and saturation regions of operation.
4. Derive an expression for the conductance in the linear region. What is meant by 'turn off'?
5. Evaluate the transconductance in the saturation region.
6. Show that a JFET having a channel heavily doped on one side behaves as a square law device.
7. What is meant by response time of a FET? Find an expression for the high frequency limit of operation.
8. How does the series resistances effect the performance of the FET in (a) linear region, (b) saturation region?

5. MOSFET

In this chapter we study an extensively used semiconductor device—Metal Oxide Semiconductor Field Effect Transistor also called the Insulated Gate Field Effect Transistor—IGFET. This transistor is easily fabricated in an integrated form and offers several advantages over other transistors. Moreover the theory of operation is of great interest to research workers as it is related to the surface properties of semiconductor surfaces. Extensive research is being done on surface properties and one of the surfaces studied extensively is the thermally oxidised silicon surface. One reason for the study of thermally oxidised surface is the ease with which it can be formed. When a semiconductor has an oxide layer formed on it the impurities are a minimum and hence less complications. We first study the surface properties of an oxidised silicon surface and later the operation of MOSFET.

5.1.1 *MOS structure*

The surface properties are normally studied on a simple structure called MOS (MIS). The structure is fabricated by forming a thin oxide layer of known thickness on a *P* or *N* type substrate. The surface of the oxide is metallised and contacts are made to the substrate and metallisation.

For theoretical study we consider an ideal MOS structure in which there are no charges present in the oxide layer. The work function differences are also neglected. The effects of charges in the oxide layer and work function differences will be considered in a later section.

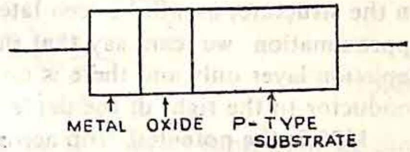


Fig. 5.1.

Under equilibrium conditions in an ideal structure the energy bands in the semiconductor are straight. This condition is referred to as a flat band condition (Fig. 5.2a).

When a potential difference is established across the structure making the metal side (gate) negative, a negative potential appears at the interface separating the oxide and semiconductor. Under the influence of the electric field set up near the surface, electrons move away from the interface increasing the hole density at the surface. This accumulation of holes is interpreted as due to the bands bending upwards; the Fermi level being straight (the Fermi level is not disturbed as there is no current flow in the semiconductor). The band diagram is given in Fig. 5.2b.

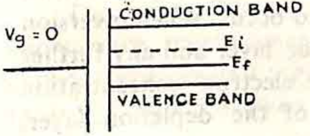


Fig. 5.2a.

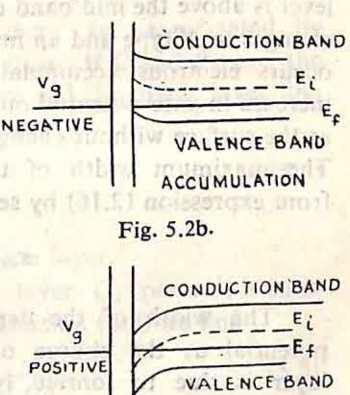


Fig. 5.2b.

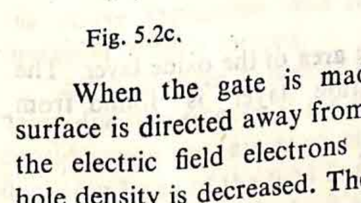


Fig. 5.2c.

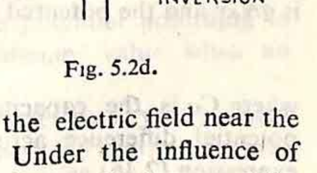


Fig. 5.2d.

When the gate is made positive, the electric field near the surface is directed away from the surface. Under the influence of the electric field electrons move towards the oxide layer and the hole density is decreased. The bands bend downward as indicated in Fig. 5.2c. It can be assumed that within a small region near the surface there are no mobile carriers and a depletion layer is formed. The depletion approximation is confirmed by experiments on the structure, as will be seen later. On the basis of the depletion approximation we can say that the electric field exists within the depletion layer only and there is no electric field in the bulk semiconductor to the right of the depletion layer.

If V_s is the potential drop across the depletion layer its width can easily be found as

$$x_d = \sqrt{\frac{2\epsilon V_s}{qN_a}} \tag{2.16}$$

As the gate voltage is increased the potential drop across the depletion layer increases, and the depletion layer widens. When the potential drop becomes equal to the Fermi potential φ_p the semiconductor at the surface becomes intrinsic. The width of the depletion layer increases till the potential drop across it becomes equal to $2\varphi_p$. In this condition the bands are bent to an extent the mid band energy level goes below the Fermi level. As the Fermi level is above the mid band energy level the semiconductor surface changes to N type and an inversion is said to occur. When inversion occurs electrons accumulate at the surface layer and any further increase in gate potential only increases the electron concentration at the surface without changing the width of the depletion layer. The maximum width of the depletion layer can easily be found from expression (2.16) by setting $V_s = 2\varphi_p$.

$$x_{d(\max)} = \sqrt{\frac{4\epsilon\varphi_p}{qN_a}} \tag{5.1}$$

The width of the depletion layer can be related to the gate potential as the charge on the semiconductor side of the oxide layer is due to ionised impurity atoms. If x is the width of the depletion layer the charge per unit area on the semiconductor side is $qN_a x$ and the potential difference across the oxide layer

$$V_O = \frac{qN_a x}{C_O}$$

where C_O is the capacitance per unit area of the oxide layer. The potential difference across the depletion layer is found from expression (2.16) as

$$V_s = \frac{qN_a x^2}{2\epsilon}$$

The gate potential

$$V_g = V_O + V_s$$

$$= \frac{qN_a x}{Co} + \frac{qN_a x^2}{2\epsilon}$$

Solving the quadratic in x we have

$$x = \frac{\epsilon}{Co}\left\{-1 + \sqrt{1 + \frac{2V_g Co^2}{\epsilon q N_a}}\right\} \tag{5.2}$$

Expanding the square root in the above expression we can approximately find the width of the depletion layer as

$$x \approx \frac{V_g Co}{qN_a} \tag{5.3}$$

The width of the depletion layer is approximately proportional to the gate potential.

The value of the gate potential for which an inversion layer is formed can be found by substituting the value of x from (5.1) in (5.2).

5.1.2 *Investigations on the MOS structure*

The properties of the MOS structure are investigated by capacitance measurements on the structure. If C and Co are the capacitances per unit area of the structure and the oxide layer, the ratio of C to Co is given by the expression

$$\frac{C}{Co} = \frac{1}{1 + \frac{Co}{C_d}} \tag{5.4}$$

where C_d is the capacitance of the depletion layer.

The capacitance of the depletion layer C_d per unit area is ϵ/x. Substituting the value of x from expression (5.2) we find

$$\frac{C}{Co} = \frac{1}{\sqrt{1 + \frac{2V_g Co^2}{\epsilon q N_a}}} \tag{5.5}$$

The ratio C/Co decreases with the gate potential according to the above expression and reaches a minimum value when an inversion layer is formed.

After the formation of the inversion layer we have a charged layer due to electrons at the interface and we have also to take into consideration the carrier generation and recombination within the depletion layer of the field induced junction. When measurements on the structure are made at a low frequency there is enough time for the inversion layer charge to follow the incremental changes in

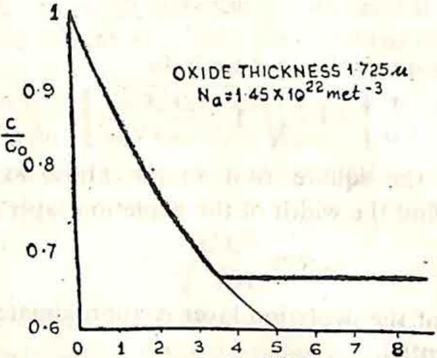


Fig. 5.3.

$N_a = 1.45 \times 10^{22}$ met^{-3}; $n_i = 1.45 \times 10^{16}$ met^{-3}
$\varphi_p = 0.3594$ V; $\varphi_{\text{inv}} = 0.719$ V; oxide thickness 1.725μ
$C_o = 2 \times 10^{-4}$ farads
$x_{d(\text{max})} = 2.532 \times 10^{-7}$ m; $V_{g(\text{inv})} = 3.66$ V

V_g(V)	C/C_o		
1	0.866	4	0.655
2	0.775	5	0.6128
3	0.707		
3.66	0.6714		

gate potential and the structure appears to have a capacitance equal to the oxide layer. The ratio C/Co increases to unity as the gate potential is raised. For silicon it is found that the above phenomenon occurs in the frequency range 5 to 100 Hz.

When measurements are made at a high frequency there is not enough time for the charges in the inversion layer to change and the structure capacitance remains constant at the minimum value

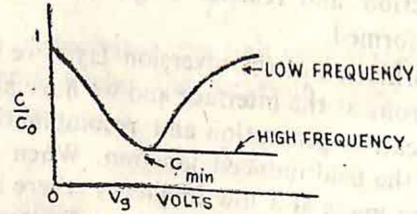


Fig. 5.4.

and is dependent on the potential. The variation of *C/Co* with frequency and gate potential are shown in Fig. 5.4.

5.1.3 *Departures from ideal*

The flat band condition assumed for the MOS structure implies that there is no potential difference across the oxide layer. This is possible only when the metal and semiconductor work functions are equal. (The work function for a material is measured from the Fermi level to the vacuum energy level which is taken as the reference.) If the work functions of the metal and semiconductor are not equal there is a potential difference across the oxide layer and the bands in the semiconductor are bent at the edge of the oxide layer.

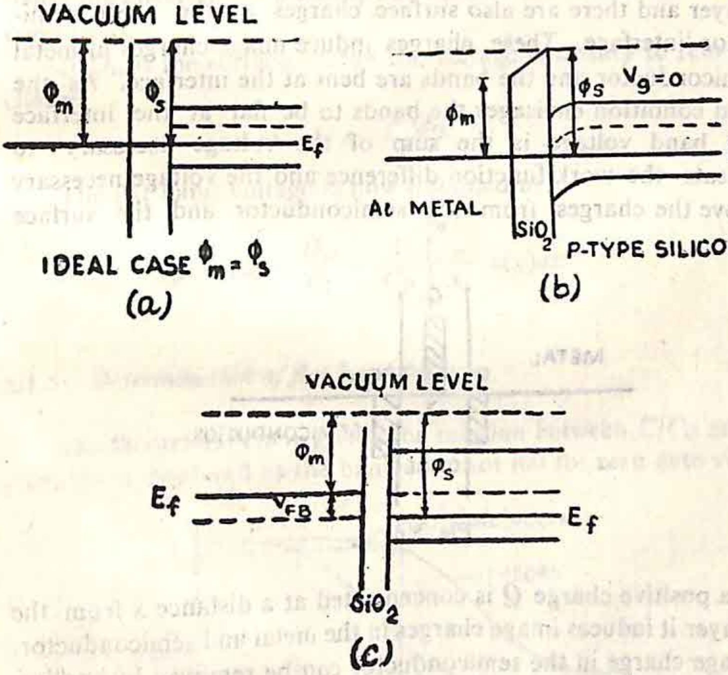


Fig. 5.5.

It is convenient to assign a flat band voltage with the structure so that the application of the voltage to the gate makes the bands flat. The flat band voltage equals the metal semiconductor work function φ_{ms}. It depends on the metal used for the gate and the substrate concentration in the semiconductor.

Value of φ_{ms}—silicon MOS structure in volts

Substrate concentration	$AlSiO_2$ —N type silicon.	$AlSiO_2$ —P type silicon	$AlSiO_2$ —N type silicon	$AlSiO_2$ —P type silicon
10^{20}	−0.36	−0.82	+0.54	+0.08
10^{21}	−0.3	−0.88	+0.6	+0.02
10^{22}	−0.24	−0.94	+0.65	−0.04
10^{23}	−0.18	−1.00	+0.72	−0.1

5.1.4 *Effect of charges*

During the process of fabrication charges are trapped in the oxide layer and there are also surface charges at the oxide semiconductor interface. These charges induce image charges in metal and semiconductor and the bands are bent at the interface. As the flat band condition envisages the bands to be flat at the interface the flat band voltage is the sum of the voltage necessary to compensate the work function difference and the voltage necessary to remove the charges from the semiconductor and the surface charge.

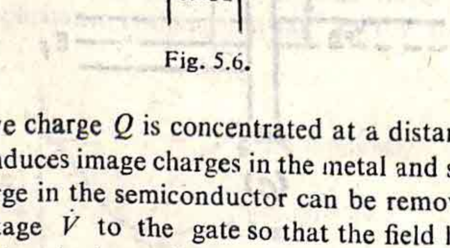


Fig. 5.6.

If a positive charge Q is concentrated at a distance x from the metal layer it induces image charges in the metal and semiconductor. The image charge in the semiconductor can be removed by application of a voltage V to the gate so that the field lines due to the fixed charge Q terminate on the metal and do not penetrate the semiconductor. The voltage V is related to the charge Q and its distance x by the expression

$$V = \frac{-xQ}{\epsilon_0} = \frac{-x}{x_0} \cdot \frac{Q}{Co} \tag{5.6}$$

If the charge is arbitrarily distributed in the oxide layer the voltage needed to remove the image charge from the semiconductor

$$V = -\frac{1}{Co}\int_0^{x_0} \frac{x}{x_0} \cdot \rho(x)dx \tag{5.7}$$

where ρ is the charge at a distance x from the gate.

In addition to charges in the insulator there is a sheet charge present at the oxide semiconductor interface. This sheet charge is due to ionic silicon entering the oxide during oxidation and annealing It is fixed and is often located at a distance of about 100 Å from the oxide semiconductor interface. It is independent of the oxide thickness and is a strong function of oxidation and annealing conditions.

If Q_{ss} is the surface charge the voltage necessary to remove the charge

$$V = \frac{-Q_{ss}}{Co} \tag{5.8}$$

The flat band voltage is now obtained as

$$V_{FB} = \varphi_m - \frac{Q_{ss}}{Co} - \frac{1}{Co}\int_0^{x_0} \frac{x}{x_0}\ \rho(x)dx \tag{5.9}$$

5.1.5 *Determination of flat band voltage*

The theoretical curve giving the relation between C/Co and gate potential is displaced as the bands are not flat for zero gate voltage.

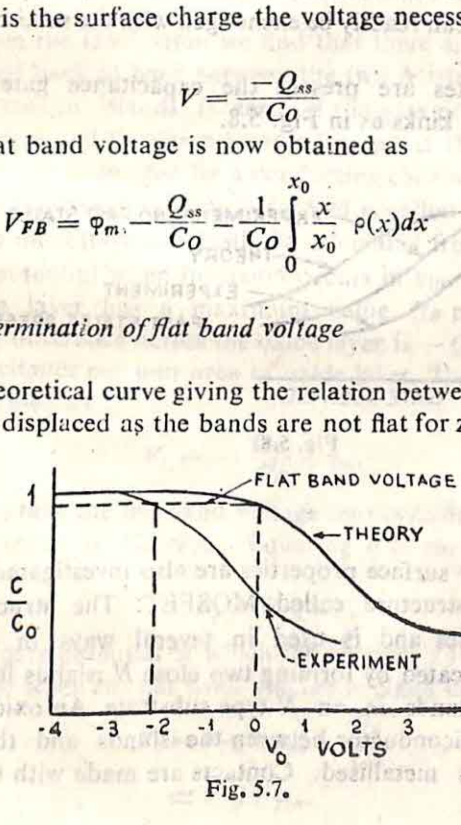


Fig. 5.7.

The experimental curve is displaced and the displacement of the experimental curve from the theoretical curve enables the determination of the flat band voltage.

5.1.6 *Effect of surface states on the capacitance—gate voltage characteristic*

We know that the bands in the semiconductor near the oxide layer are bent when the gate potential is changed. If there are energy states within the forbidden gap the probability of their occupation depends on the energy separation between them and the Fermi level which remains constant. When the bands are moved up the energy difference increases resulting in the states becoming vacant. When the bands are moved down the states are filled with electrons. Thus we see that there is a charge exchange between the surface states and the semiconductor when the bands are bent. The states whose charge can readily be exchanged with the semiconductor are called fast states.

When fast states are present the capacitance gate voltage characteristic shows kinks as in Fig. 5.8.

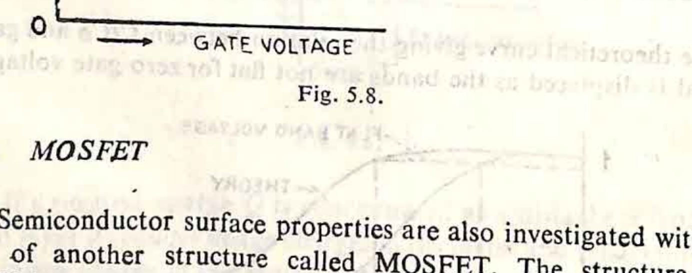


Fig. 5.8.

5.2.1 *MOSFET*

Semiconductor surface properties are also investigated with the help of another structure called MOSFET. The structure has amplifying properties and is used in several ways in electronic circuitry. It is fabricated by forming two close *N* islands in a *P*-type substrate or *P* islands in an *N*-type substrate. An oxide layer is formed on the semiconductor between the islands and the top of the oxide layer is metallised. Contacts are made with the two *N* islands and the metallisation. One of the N islands is called the source, the other the drain, and the metallisation the gate. One obvious advantage of the fabrication is that no isolation is needed between units.

5.2.2 *Channel formation—turn on voltage—channel conductance*

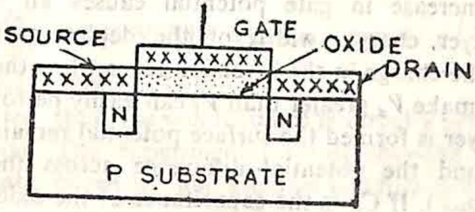


Fig. 5.9.

From the fabrication we find that there are two PN junctions connected back to back between the two N islands and the conductance between islands is zero. If the gate potential is made sufficiently large and inversion layer is formed at the surface, the two N islands are connected by a conducting channel. The gate potential at which the inversion layer is formed is called the turn on voltage. The turn on voltage can easily be calculated from the fact that the surface potential when inversion occurs in φ_{inv}. The charge in the depletion layer has a maximum value Q_B per unit area and the potential difference across the oxide layer is $-Q_B/C_O$, where C_O is the capacitance per unit area of oxide layer. The turn on voltage is now obtained as

$$V_t = -\frac{Q_B}{C_O} + \varphi_{\text{inv}} \tag{5.10}$$

If we take the flat band voltage into consideration the effective gate potential is $V_g - V_{FB}$. Equating it to the turn on voltage we have

$$V_g = V_{FB} + V_t \tag{5.11}$$

The gate potential at which the channel is formed or the turn on voltage when the flat band voltage is taken into consideration is

$$V_t = V_{FB} + \varphi_{\text{inv}} - \frac{Q_B}{C_O} \tag{5.12}$$

$$\approx V_{FB} + \varphi_{\text{inv}} \tag{5.12 a}$$

The turn on voltage can be written as $V_{FB}+\varphi_{\text{inv}}$ when $\frac{Q_B}{Co}$ can be neglected.

If the flat band voltage is negative and greater than φ_{inv} a channel is formed below the oxide layer even when the gate is at zero potential.

It was found earlier in this chapter that once the inversion layer is formed increase in gate potential causes an increase in the inversion layer charge, width of the depletion layer remaining constant. The charge in the inversion layer when the gate potential is raised to make V_g greater than V_t can easily be found. Once the inversion layer is formed the surface potential remains at a constant value φ_{inv} and the potential difference across the oxide layer is $(V_g-V_{FB}-\varphi_{\text{inv}})$. If Co is the capacitance of the oxide layer per unit area the charge on the semiconductor side of the oxide layer per unit area is $-(V_g-V_{FB}-\varphi_{\text{inv}})Co$. This charge is the sum of the charge in the inversion layer and the charge due to ionised impurity atoms in the depletion layer. If Q_n is the charge per unit area in the inversion layer we can write

$$Q_n=-(V_g-V_{FB}-\varphi_{\text{inv}})\,Co-Q_B \tag{5.13}$$

The conductance of the channel formed between the source and drain can be easily found. If it is assumed that the potential difference between the two N islands is small the conductance of the channel

$$g=\frac{Z}{L}\int_0^{x_i}\sigma(x)dx$$

where $\sigma(x)$ is the conductance of the channel at a distance x from the oxide layer.

$$\sigma(x)=q\mu_n n(x)$$

where $n(x)$ is the density of electrons at a distance x from the oxide layer.

$$g=\frac{Z}{L}\int_0^{x_i}q\mu_n n(x)dx$$

$$=\frac{-Z}{L}\cdot Q_n \tag{5.14}$$

Substituting the value of Q_n from expression (5.13) in (5.14) we have, after using expression in (5.12)

$$g=\frac{Z\mu_n C_O}{L}[V_g-V_t] \qquad (5.15)$$

Consider a MOS structure with a substrate doping 1.45×10^{22} met^{-3}. Let the oxide layer be 1000 Å thick. Under these conditions $C_O=3.45\times10^{-4}$ farads, $\varphi_{\text{inv}}=0.718$ V, $Q_B=-5.873\times10^{-4}$ coulombs. $Q_B/C_O=-1.701$ V. The work function difference for AlSiO$_2$ *P*-type silicon is -0.94 V. Therefore the turn on voltage $=V_{FB}+\varphi_{\text{inv}}\frac{Q_B}{C_O}=2.46$ V. The relation between $g\Big/\frac{Z\mu_n C_O}{L}$ and V_g-V_t is plotted in Fig. 5.10.

We thus see that the conductance is directly proportional to the gate potential. The experimental result showing the relation between the gate potential and channel conductance is a straight line graph.

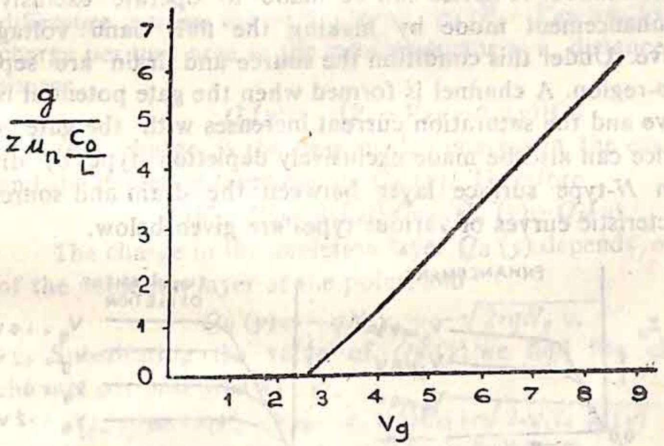


Fig. 5.10.

Often during investigations on the structure the conductance becomes zero for a negative value of the gate potential. This happens because the flat band voltage is negative and has a large value compared to the turn on voltage V_t.

Principle of operation of MOSFET

It was stated in the last section that a conducting channel exists between the source and drain even when the gate potential is zero if the flat band voltage is negative and larger than the inversion voltage. If the drain is made positive relative to the source a current flows through the channel and the channel current increases with the drain voltage initially. However, with the rise in drain potential there is a rise in channel potential and the channel pinches off (disappears) when the drain potential becomes equal to the flat band voltage. When the channel pinches off the channel current reaches a maximum or saturation value. The saturation current depends on the voltage at which the pinch off occurs and also the conductance of the channel. If the gate potential is increased the pinch off occurs for a larger drain voltage and the channel conductance is also reduced so the saturation current increases with the gate potential and the device is said to operate in the enhancement mode. Conversely if the gate is made negative the saturation current decreases and the device operates in the depletion mode. A device can be made to operate exclusively in the enhancement mode by making the flat band voltage less negative. Under this condition the source and drain are separated by a *p*-region. A channel is formed when the gate potential is made positive and the saturation current increases with the gate voltage. A device can also be made exclusively depletion type by diffusing a thin *N*-type surface layer between the drain and source. The characteristic curves of various types are given below.

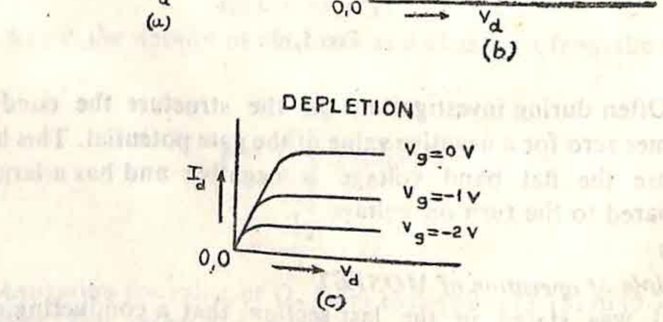


Fig. 5.11.

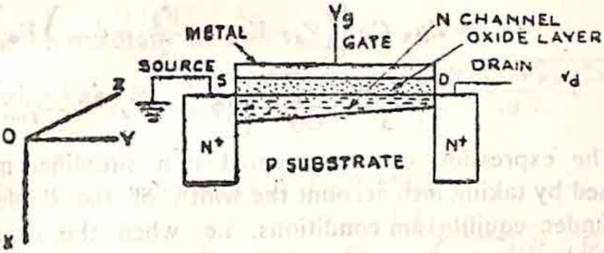


Fig. 5.12.

Current voltage relation

Consider a MOSFET having a channel of length L and width Z. Let a current I_d flow through the channel when the drain potential is V_d and gate potential V_g. If V_y is the potential in the channel at a distance y from the source the surface potential $\varphi_s(y)$ at the point is $(\varphi_{\text{inv}}+V_y)$ as the channel is formed. The potential difference across the oxide layer is $[V_g-V_{FB}-\varphi_s(y)]$ and the charge per unit area in the semiconductor at a distance y from the source

$$Q_s(y)=-[V_g-V_{FB}-\varphi_s(y)]Co$$

This charge is the sum of the charges in the channel $Q_n(y)$ and the depletion layer charge $Q_B(y)$. Therefore

$$-[V_g-V_{FB}-\varphi_s(y)]\,Co=Q_n(y)+Q_B(y)$$

The charge in the depletion layer $Q_B(y)$ depends on the width of the depletion layer at the point, and

$$Q_B(y)=-qN_a x_d=-\sqrt{2\epsilon qN_a\varphi_s}$$

Substituting the value of $Q_B(y)$ we find the charge in the channel per unit area as

$$Q_n(y)=-\{V_g-V_{FB}-\varphi_s(y)\}Co+\sqrt{2\epsilon qN_a\varphi_s(y)} \qquad (5.16)$$

The resistance of the channel for a length dy

$$dR=\frac{dy}{Z\mu_n Q_n(y)}$$

and the potential drop across a length dy of the channel

$$dV=-I_d dR$$

Rearranging, after substituting the value of dR and $Q_n(y)$, we have

$$I_d dy=Z\mu_n[V_g-V_{FB}-\varphi_s(y)]\,Co\,dV-\sqrt{2\epsilon qN_a\varphi_s(y)}\cdot dV$$

Integrating along the length of the channel and rearranging, we find

$$I_dL = Z\mu_n C_O\left\{\left(V_g - V_{FB} - \frac{V_d}{2} - \phi_{\text{inv}}\right)V_d - \frac{2}{3}\frac{\sqrt{2\epsilon qN_a}}{C_O}\left[(\varphi_{\text{inv}}+V_d)^{3/2} - \varphi_{\text{inv}}^{3/2}\right]\right. \qquad (5.17)$$

The expression can be presented in a simplified manner and analysed by taking into account the width of the depletion layer x_{d0}, under equilibrium conditions, i.e. when the drain and gate potentials are zero

$$x_{d0} = \sqrt{\frac{2\epsilon\varphi_{\text{inv}}}{qN_a}}$$

The capacitance per unit area of the oxide layer

$$C_O = \frac{\epsilon_0}{x_0} \qquad (5.17)$$

x_0 is the thickness of oxide layer and ϵ_0 its permittivity.

We find, after substituting the values of x_{d0} and C_O, that

$$\frac{\sqrt{2\epsilon qN_a}}{C_O} = \frac{2\epsilon}{\epsilon_0}\cdot\frac{x_0}{x_{d0}}\sqrt{\phi_{\text{inv}}}$$

$$= K\sqrt{\phi_{\text{inv}}} \qquad (5.18)$$

where $$K = \frac{2\epsilon}{\epsilon_0}\cdot\frac{x_0}{x_{d0}}$$

The value of K depends on the thickness of the oxide layer and the doping density in the substrate. If the doping density is low and the oxide layer thin, K becomes very small and can be neglected as an approximation.

Substituting the value of K from expression (5.18) in expression (5.17) we find

$$I_dL = Z\mu_n C_O\left\{\left(V_g - V_{FB} - \varphi_{\text{inv}} - \frac{V_d}{2}\right)V_d - \frac{2}{3}K\varphi_{\text{inv}}^{1/2}\left[(\varphi_{\text{inv}}+V_d)^{3/2} - \varphi_{\text{inv}}^{3/2}\right]\right\} \qquad (5.19)$$

The channel current I_d increases with the drain potential V_d according to the above expression till the channel disappears. The drain potential at which the channel disappears $V_{d(\text{sat})}$ and the saturation value of the current $I_{d(\text{sat})}$ can also be found.

When the channel disappears the surface potential $\varphi_s = V_{d(\text{sat})} + \varphi_{\text{inv}}$ and as the charge in the channel per unit area is zero, we have from expression (5.16)

$$0 = -[V_g - V_{FB} - \varphi_s(y)]C_O + \sqrt{2\epsilon qN_a\varphi_s(y)}$$

Writing $\varphi_s(y) = V_{d(\text{sat})} + \varphi_{\text{inv}}$

and solving the quadratic we find $V_{d(\text{sat})}$ as

$$V_{d(\text{sat})} = \left(V_g - V_{FB} - \varphi_{\text{inv}} \right) + \frac{\epsilon q N_a}{Co^2}\left[1 - \sqrt{1 + \frac{2(V_g - V_{FB})Co^2}{\epsilon q N_a}}\right] \qquad (5.20)$$

Substituting the value of $K\varphi_{\text{inv}}^{1/2}$ for $\dfrac{\sqrt{2\epsilon q N_a}}{Co}$

$$V_{d(\text{sat})} = \left(V_g - V_{FB} - \varphi_{\text{inv}} \right) + \frac{K^2\varphi_{\text{inv}}}{2}\left[1 + \sqrt{1 + \frac{4(V_g - V_{FB})}{K^2\varphi_{\text{inv}}}}\right] \qquad (5.21)$$

When K is small enough to neglect the factor in K^2, we have

$$V_{d(\text{sat})} = (V_g - V_{FB} - \varphi_{\text{inv}})$$

Using expression (5.12a)

$$V_{d(\text{sat})} = (V_g - V_t) \qquad (5.22)$$

Substituting the value of $V_{d(\text{sat})}$ from expression (5.22) in (5.19) the saturation value of the drain current is found as

$$I_{d(\text{sat})} = \frac{Z\mu_n Co}{2L}(V_g - V_t)^2 \qquad (5.23)$$

The device works as a square law device.

The theoretical characteristics of a MOSFET are given in Fig. 5.13.

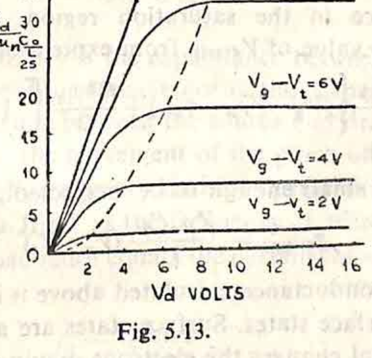


Fig. 5.13.

Conductance and transconductance

The two regions of operation of the device are: (a) linear region, (b) the saturation region.

In the linear region of operation the conductance of the channel

$$g = \left.\frac{\partial I_d}{\partial V_d}\right|_{V_g \text{ const.}} \tag{5.24}$$

By assuming that $V_d \ll \varphi_{\text{inv}}$ we can expand the bracketed term in (5.19) and find that

$$I_d L = Z\mu_n Co\left\{\left(V_g - V_{FB} - \varphi_{\text{inv}} - \frac{V_d}{2}\right)V_d - K\varphi_{\text{inv}} V_d\right\}$$

$$K\varphi_{\text{inv}} = -\frac{Q_B}{Co}$$

The conductance in the linear region

$$g = \frac{Z\mu_n Co}{L}(V_g - V_t) \tag{5.25}$$

The is the same as (5.15).

The conductance in the saturation region is almost zero as the channel disappears under saturation conditions.

The transconductance of the device

$$g_m = \left.\frac{\partial I_d}{\partial V_g}\right|_{V_d \text{ const.}}$$

From expression (5.17) we find that the transconductance

$$g_m = \frac{Z\mu_n Co}{L} \cdot V_d \tag{5.26}$$

The above expression gives the transconductance in the linear region of operation. It is proportional to the drain voltage. The transconductance in the saturation region can be found by substituting the value of $V_{d(\text{sat})}$ from expression (5.21) in (5.26)

$$g_{m(\text{sat})} = \frac{Z\mu_n Co}{L}\left\{(V_g - V_{FB} - \varphi_{\text{inv}}) + \frac{K^2\varphi_{\text{inv}}}{2}\left[1 + \sqrt{1 + \frac{4(V_g - V_{FB})}{K^2\varphi_{\text{inv}}}}\right]\right. \tag{5.27}$$

When K is small enough to be neglected

$$g_{m(\text{sat})} = \frac{Z\mu_n Co}{L}(V_g - V_t) \tag{5.28}$$

The transconductance calculated above is in an ideal case where there are no surface states. Surface states are always present, when the gate potential changes the electrons moving into the conducting channel may occupy the surface states and hence become unavailable for conduction. This results in a decrease in transconductance. However, if the frequency of operation is high there may not be enough time for the electrons to move into the surface states and the transconductance approaches the value in expression (5.28).

Applications of MOSFET

A MOSFET is used as a resistor, an amplifying device and as a memory element.

When the device is used in the linear region it functions as a resistor. As the conductance depends on the gate potential the resistance value can be changed by biasing.

The MOSFET is used as an amplifying device in the saturation region with the source common to the input and output. The input resistance is very high—of the order of 10^{14} ohms, it is much higher than the input resistance of the junction field effect transistor. Moreover the input resistance is independent of temperature. The input resistance of a JFET changes with temperature as the reverse saturation current varies. The high input resistance is an advantage in circuits where loading is to be avoided. The equivalent circuit of of the MOSFET is given below.

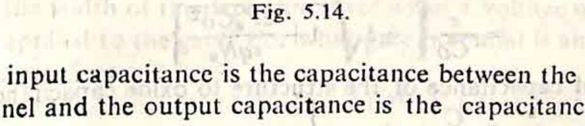


Fig. 5.14.

The input capacitance is the capacitance between the gate and the channel and the output capacitance is the capacitance of two *PN* junctions appearing between the source and drain. The output conductance is due to the movement of the pinch off point and is very small. The high frequency limit of operation is found in the same manner as for a JFET as the frequency at which the current through the input capacitance equals the current through the current generator

$$f_0 = \frac{g_m}{2\pi C_{in}}$$

$$C_{in} = CoZL \text{ and } g_m = \frac{Z\mu_n Co}{L} \cdot V_d$$

$$f_0 \approx \frac{\mu_n \cdot V_d}{L^2} \tag{5.29}$$

The transit time of carriers through the channel

$$t = \frac{L}{v} \approx \frac{L}{\mu E}$$

where E is the electric field intensity in the channel and is equal to $\frac{V_d}{L}$. Substituting the value of E we find

$$t \sim \frac{L^2}{V_d} \tag{5.30}$$

The high input resistance of the MOSFET is made use of in utilising the MOSFET as a dynamic memory element. When a pulse is applied to the input terminals the input capacitance is charged and the voltage is retained on the gate terminal as the leakage resistance is very small.

Noise

The sources of noise in a MOSFET are:

1. Thermal noise due to current charge carriers in the channel.
2. Generation recombination.
3. Noise due to random fluctuations of carriers in the surface states.

Important Formulae

Width of the depletion layer in a MOS structure

$$= \frac{\epsilon}{Co}\left\{ +1\sqrt{1+\frac{2V_g Co^2}{\epsilon q N_a}} \right\}$$

Ratio of capacitance of the structure to oxide capacitance

$$\frac{C}{Co} = \frac{1}{\sqrt{1+\frac{2V_g Co^2}{\epsilon q N_a}}}$$

Flat band voltage $V_{FB} = \varphi_{ms} - \frac{Q_{ss}}{Co} - \frac{1}{Co}\int_0^{x_0} \frac{x}{x_0}\,\rho(x)dx$

Turn on voltage $V_t = V_{FB} + \varphi_{\text{inv}} - \frac{Q_B}{Co}$

Channel current $$I_d = \frac{Z\mu_n Co}{L}\left\{\left(V_g - V_{FB} - \varphi_{\text{inv}} - \frac{V_d}{2}\right)V_d - \frac{2}{3}\sqrt{\frac{2\epsilon q N_a}{Co}}\left[(\varphi_{\text{inv}} + V_d)^{3/2} - V_d^{3/2}\right]\right\}$$

$$V_{d(\text{sat})} = (V_g - V_{FB} - \varphi_{\text{inv}}) + \frac{K^2\varphi_{\text{inv}}}{2}\left[1 + \sqrt{1 + \frac{4(V_g - V_{FB})}{K^2\varphi_{\text{inv}}}}\right]$$

$$I_{d(\text{sat})} = \frac{Z\mu_n Co}{2L}(V_g - V_t)^2$$

$$f_0 \approx \frac{\mu_n V_d}{L^2}$$

Useful Data

n_i for silicon $= 1.45 \times 10^{16}$ met^{-3}

$\epsilon_0 = 8.854 \times 10^{-12}$ farad met

ϵ_r (silicon) $= 11.7$ ϵ_r (oxide) $= 3.9$

Questions

1. Briefly discuss the shape of the bands at the semiconductor surface in a MOS structure when a voltage is applied to the gate. What is inversion? When will it occur?
2. Derive the relation between the gate potential and the width of the depletion layer.
3. An oxide layer of 2000 Å is formed on a *P*-type silicon substrate with a doping density 1.45×10^{20} met^{-3}. What is the width of the depletion layer when a voltage of one V is applied to the gate? For what gate potential is an inversion layer formed?
4. Evaluate the ratio of the MOS structure capacitance to the oxide layer capacitance. When will the value be minimum?
5. Find the minimum value of the ratio C/C_0 for the structure in Question 3.
6. What is flat band voltage? Determine its value for a MOS structure. How can it be determined experimentally?
7. Evaluate the flat band voltage for $AlSiO_2 - P$-type silicon structure with an oxide layer thickness 1500 Å if the

sheet charge has a value 5×10^{-4} coulombs met^{-2}. The substrate doping is 1.45×10^{21} met^{-3} (neglect charge in oxide layer).

8. Distinguish between: Enhancement type, Depletion type and Enhancement depletion type of MOSFETs. Draw typical characteristic curves.
9. Derive the voltage current relation for a MOSFET.
10. When will the current through a MOSFET saturate? Evaluate the drain voltage when the current saturates.
11. Derive expressions for: (a) conductance in the linear region of operation, (b) mutual conductance in the saturation region.
12. Show that the device functions as a square law device, when the substrate doping is low and oxide layer thin.
13. Discuss the effects of surface states on the mutual conductance.
14. Draw the equivalent circuit of a MOSFET.
15. Derive an expression for the high frequency limit of operation of the device.

6. Microwave Semiconductor Devices

In this chapter we study some of the semiconductor devices used in the microwave frequency range. The diodes and transistors discussed in the earlier chapters have cut off frequencies much below the microwave frequency range which is normally above 100 MHz. The advantages of non-linearity in the operation of the devices and negative resistance offered by devices under certain conditions are made use of in the operation of the devices in the microwave region. The devices studied in this chapter are varactor, *PIN* diode, tunnel diode and the devices based on the Gunn effect. As the theories are fairly advanced the basic principles of operation are only stressed.

6.1 Parametric Amplification

An active device—vacuum tube or transistor—amplifies a signal by transferring energy from the source of steady current to the signal. Energy can also be transferred from a strong signal called pump signal to a weak signal through a passive non-linear circuit element. The amplification produced in this manner is called parametric amplification.

The parametric amplification can easily be illustrated by an imaginary experiment. Consider a resonant A.C. circuit in which oscillations exist. During the course of oscillations energy oscillates between the electric field in the capacitor and the magnetic field in the coil. The oscillations continue indefinitely if there are no losses.

The amplitude of oscillations can be increased by putting in energy into the signal at the proper moment. One of the ways of putting in energy into the signal is by increasing the separation between the capacitor plates when the voltage between the plates is maximum. Some work is done in increasing the separation between the plates and the work done gets converted into signal energy. The capacitor plates should be brought closer when the voltage between them is zero, otherwise energy flows away from the signal. Separating the plates of the capacitor mechanically at the proper time is not possible, as such one has to use a capacitor whose capacitance varies with voltage. It has been shown that the capacitance of a reverse biased junction is a function of the voltage across it and the variation of capacitance can be conveniently used to transfer energy to a signal. The alternating voltage used to vary the capacitance is called the pump signal.

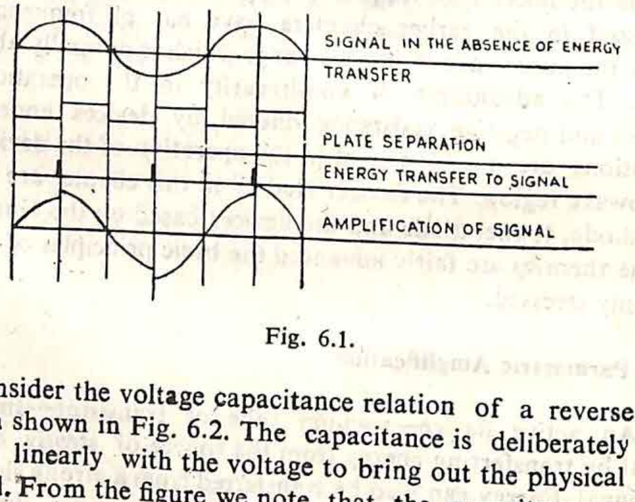

SIGNAL IN THE ABSENCE OF ENERGY TRANSFER
PLATE SEPARATION
ENERGY TRANSFER TO SIGNAL
AMPLIFICATION OF SIGNAL

Fig. 6.1.

Consider the voltage capacitance relation of a reverse biased junction shown in Fig. 6.2. The capacitance is deliberately shown to vary linearly with the voltage to bring out the physical process involved. From the figure we note that the capacitance reaches a minimum value equivalent to an increase in plate separation when the voltage is minimum, i.e. once in a cycle. It has been shown earlier that energy is pumped into the signal when the separation is increased when the voltage across the capacitor due to the signal is maximum. Therefore we can easily see that the pump frequency should be twice the signal frequency.

We can mathematically show that energy flows into the signal from the pump. Let the capacitance of a reverse biased junction be given by the expression $C=KV^{-m}$, where V is the voltage

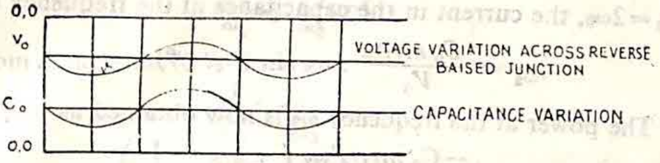


Fig. 6.2.

across the junction and the exponent m depending on the nature of the junction. If the voltage V is the sum of steady component V_0 and an alternating component v we have

$$C=K(V_0+v)^{-m}$$

$$=KV^{-m}\left(1+\frac{v}{V_0}\right)^{-m}$$

When v is small compared to V_0 the above expression can be approximated and written as

$$C=KV^{-m}\left(1-m\cdot\frac{v}{V_0}\right)$$

$$=C_0\left(1-\frac{mv}{V_0}\right) \qquad (6.1)$$

If the alternating component of the current is due to a signal and a pump voltage we have

$$v=v_1\sin\omega_1 t+v_2\sin(\omega_2 t+\theta)$$

where v_1 and v_2 are the maximum values of the pump and signal voltages whose frequencies are ω_1 and ω_2. θ is the phase difference between the signal and pump voltages.

The current through the capacitor

$$i=\frac{dQ}{dt}=\frac{dQ}{dv}\cdot\frac{dv}{dt}$$

$$=C_0\left(1-\frac{mv}{V_0}\right)\frac{dv}{dt} \qquad (6.2)$$

The factor $C_0\,\frac{dv}{dt}$ contains currents at frequencies ω_1 and ω_2, while the factor $C_0\,\frac{mv}{V_0}\cdot\frac{dv}{dt}$ contains currents at side-band frequencies $(\omega_1+\omega_2)$ and $(\omega_1-\omega_2)$. The current at the side band frequency $(\omega_1-\omega_2)$ can easily be found as

$$i_{(\omega_1-\omega_2)}=\frac{C_0\,mv_1v_2\,(\omega_1-\omega_2)}{V_0}\sin\overline{(\omega_1-\omega_2)\,t-\theta} \qquad (6.3)$$

If $\omega_1=2\omega_2$, the current in the capacitance at the frequency ω_2

$$i_{\omega_2}=\frac{C_0\,mv_1v_2}{V_0}\cdot\omega_2\sin(\omega_2t-\theta) \tag{6.4}$$

The power at the frequency ω_2 is now obtained as

$$P_{\omega_2}=\frac{-C_0\,mv_1v_2^2\omega_2}{V_0}\left(\sin^2\theta-\frac{1}{2}\right) \tag{6.5}$$

The power at the difference frequency ω_2 is negative indicating that power flows out of capacitor, i.e. energy is transferred to the signal. Maximum power transfer occurs when $\theta=\pi/2$.

Manley Rowe equations

We have till now studied a special case where the pump frequency is double the signal frequency. In a general case where two alternating voltages of frequencies ω_1 and ω_2 are applied to a non-linear capacitor a number of side-band frequencies are generated by the capacitor, the non-linear capacitor functioning as a mixer. During the process of mixing there is energy exchange between the applied signals and side bands. The power relations existing between the signals and side bands are given by Manley Rowe equations as

$$\sum_{n=-\infty}^{n=\infty}\sum_{m=0}^{m=\infty}\frac{mP_{mn}}{m\omega_1+n\omega_2}=0 \tag{6.6}$$

$$\sum_{m=-\infty}^{m=\infty}\sum_{n=0}^{n=\infty}\frac{nP_{mn}}{m\omega_1+n\omega_2}=0 \tag{6.7}$$

P_{mn} is the power input to the capacitor at a frequency $m\omega_1+n\omega_2$.

Frequency selective networks are used so select the appropriate side band. The result of mixing and power exchange can be illustrated by taking up certain special cases.

Non-inverting case

Consider a case where a frequency $\omega_1+\omega_2$ is selected and all other side band frequencies rejected. We have from (6.6)

$$\frac{P_1}{\omega_1}+\frac{P_{11}}{\omega_1+\omega_2}=0$$

P_{11} is the power at the frequency $\omega_1+\omega_2$. Writing P_{11} as P_3 and $\omega_1+\omega_2$ as ω_3

$$\frac{P_1}{\omega_1}+\frac{P_3}{\omega_3}=0 \tag{6.8}$$

From expression (6.7) we have

$$\frac{P_2}{\omega_2}+\frac{P_3}{\omega_3}=0 \tag{6.9}$$

If power is fed to the non-linear capacitor at the frequency ω_1, we find that P_2 is positive and P_3 negative. This implies a power input at frequencies ω_1 and ω_2 and power output from the capacitance at a frequency ω_3. The amplifier is called an up convertor or modulator as there is power in the upper side band. The power gain

$$G_{13}=\frac{P_3}{P_1}=\frac{\omega_3}{\omega_1} \tag{6.10}$$

Inverting case

If the frequency selecting network is adjusted for the lower side band $\omega_2-\omega_1$ we have from expressions (6.6) and (6.7)

$$\frac{P_1}{\omega_1}-\frac{P_3}{\omega_2-\omega_1}=0 \tag{6.11}$$

$$\frac{P_2}{\omega_2}+\frac{P_3}{\omega_2-\omega_1}=0 \tag{6.12}$$

When power is fed to the capacitor at the frequency ω_2 called the pump frequency we find from expression (6.12) that P_3 is negative, i.e. power is obtained from the capacitor at the frequency $(\omega_2-\omega_1)$. As P_3 is negative we find from equation (6.11) that P_1 is also negative. We thus see that there is power transfer from the pump frequency ω_2 to frequencies ω_1 and ω_3. If the power P_1 is due to a signal it is obvious that the signal is amplified as there is power transfer into the signal. If ω_2 is chosen equal to $2\,\omega_1$ (this case has been discussed earlier) large amount of power is obtained at the frequency ω_1 but it is not possible to separate out the amplified signal of frequency ω_1 from the signal of frequency $\omega_2-\omega_1$. The amplifier is called a degenerate amplifier. If $\omega_2-\omega_1$ is different from ω_1 power is dissipated at this frequency and the amplifier is called a non-degenerate amplifier. The frequency $(\omega_2-\omega_1)$, is called the idler frequency and its existence is essential to maintain the necessary phase difference between signals for the necessary power transfer.

6.2.1 *Varactor*

Varactor is a junction diode whose capacitance is a function of the applied voltage, hence can be used as a non-linear device. It was found in Chapter 2 that the depletion layer capacitance varies inversely as the square root of the applied voltage for an abrupt junction and inversely as the cube root of the applied voltage for a graded junction. The capacitance variation with voltage is dependent on the nature of the junction.

We now consider the variation of capacitance with voltage for an arbitrary distribution of impurities. Consider a P^+N junction with the donor density on the N side varying as Bx^m where B and m are constants and x—the distance from the metallurgical junction on the N side.

The Poisson's equation gives

$$\frac{\partial^2 V}{\partial x^2} = -\frac{\rho}{\epsilon} = -\frac{qBx^m}{\epsilon} \tag{6.13}$$

Integrating it and making use of the fact that there is no electric field in the bulk semiconductor $\frac{dV}{dx}=0$ when $x=W$, we find the width of the depletion layer for a voltage V existing across the layer as

$$W = \left[\frac{(m+2)\epsilon V}{qB}\right]^{1/(m+2)} \tag{6.14}$$

The capacitance of the junction

$$C = \frac{A\epsilon}{W} \propto AV^{-1/(m+2)} \tag{6.15}$$

From the above we find that $C \propto V^{-1/2}$, where $m=0$ (abrupt junction) and $C \propto V^{-1/2}$, where $m=1$ (graded junction).

If the doping profile on the N side is such as to make $m=-3/2$, the capacitance $C \propto V^{2}$

$$\tag{6.16}$$

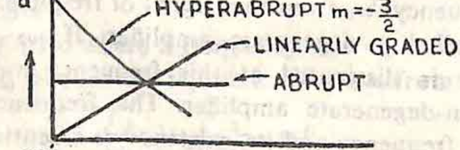


Fig. 6.3.

The doping profile making $m=-3/2$ can be obtained by epitaxial process and the junction is referred to as a hyper abrupt junction.

When the capacitance of the hyper abrupt forms a part of a resonant circuit the resonant frequency

$$\omega = \frac{1}{\sqrt{LC}} \propto V \tag{6.17}$$

The resonant frequency is proportional to the voltage, a factor which is very useful.

Figures of merit of a varactor diode

As stated earlier the varactor diode is used as a non-linear element in microwave circuits as such the most important requisite is the variation of capacitance with voltage. The capacitance is maximum under unbiased conditions and is a minimum when the reverse voltage is maximum. The maximum reverse voltage is naturally the breakdown voltage of the junction. The ratio of maximum capacitance to the minimum capacitance is one of the figures of merit of the varactor. The ratio is large when the junction is hyperabrupt.

The physical property made use of is the variation of capacitance, presence of series resistance is parasitic and is a source of power wastage. The series resistance is due to the semiconductor material on either side of the junction and the body resistance when the diode is manufactured by an epitaxial process. Moreover when the bias changes the width of the depletion layer changes and the series resistance is altered due to changes in the unswept portion of the material. This incremental resistance is large in hyper abrupt junctions. In addition to the series resistance one has to take into consideration the resistance parallel to the junction arising out of carrier generation and recombination within the depletion layer.

The maximum frequency of operation or the cut off frequency depends on the series resistance and the junction capacitance and is given by the expression

$$\omega_{c0} = \frac{1}{C_j R_s} \tag{6.18}$$

where C_j is the junction capacitance and R_s the series resistance.

The cut off frequency depends on the bias as the capacitance is voltage dependent. The frequency difference between bias extremes is referred to as the dynamic cut off frequency.

$$\omega_c = \frac{1}{R_s}\left(\frac{1}{C_{j\,\min}} - \frac{1}{C_{j\,\max}}\right) \tag{6.19}$$

The ratio of the energy stood in the capacitor to the energy dissipated in the series resistance is referred to as the dynamic quality factor and is given by the expression

$$Q = \frac{1}{\omega R_s} \tag{6.20}$$

The equivalent circuit of the varactor

The equivalent circuit of the varactor useful for investigating its performance when used in circuits is given below.

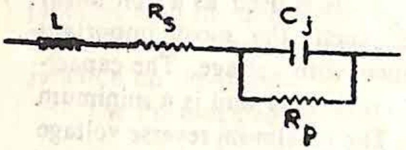


Fig. 6.4.

R_s is the series resistance comprising the resistance of the semiconductor regions on either side of the junction and the resistance at the contacts. L is the lead inductance. R_p is the leakage resistance of the diode. These components of the equivalent circuits are parasitic as they do not contribute to the physical operation of the device and attempts are made to reduce them.

6.3 Snap Back Diode

The variation in the capacitance of a diode can be maximised by changing the diode from forward bias to reverse bias. When a diode is forward biased the junction has storage capacitance due to the large concentration of minority carriers on either side of the depletion layer. The stored charge can be reclaimed when the diode is suddenly switched from a forward bias to a reverse bias. It was stated in Chapter 2 that when the voltage across a diode is reversed suddenly from forward conduction a current continues to flow till all the stored charge is cleared and the time is referred to as the storage time. It is not a desirable characteristic when the diode is used for switching. However the large reverse current transient is rich in harmonics and is used for harmonic generation at microwave frequencies. When the reverse transient is made use of, it is necessary to increase the reverse transient and provide conditions for a rapid discharge.

We know that under forward biased conditions minority carrier injection occurs across the junction and there is diffusion and recombination of the minority carriers. The recombination is reduced to a minimum by using material with long lifetimes and the diffusion is inhibited by generating a built-in electric field. The impurity doping is as indicated in Fig. 6.5.

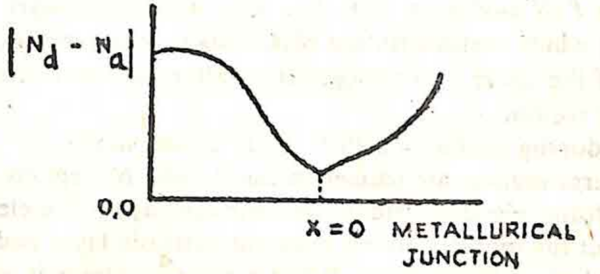


Fig. 6.5.

The impurity density on either side of the metallurgical junction is such as to oppose diffusion of minority carriers. The minority carriers are therefore concentrated on either side of the depletion layer. On reversal of voltage the stored charge is reclaimed in a very short time. The behaviour of the snap back diode is best illustrated by the current waveform generated when a large sinusoidal voltage is applied.

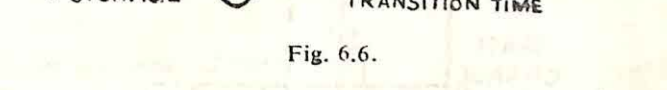


Fig. 6.6.

6.4.1 *PIN diode*

The reactance of a reverse biased junction diode is small when operated at microwave frequencies, as such it cannot be used in the microwave frequency region. The formation of an intrinsic layer between two heavily doped P and N regions reduces the capacitance of the reverse biased junction considerably and the resultant diode referred to as *PIN* diode is used as a switch in the microwave region. The intrinsic layer is about 50μ thick. It is not technologically possible to form an intrinsic region between two heavily doped *P* and *N* regions. Normally a high resistivity *P* layer or an *N* layer is formed between two heavily doped regions. The structures formed are called *PπN* and *PνN* diodes. Another interesting property of *PπN* or *PνN* diodes is that the high resistivity layer offers a resistance which varies with the bias. This property makes possible the use of the diode as a variolosser or attenuator in the microwave frequency region.

The doping profile in a *PIN* diode is as shown in Fig. 6.7. Space charge regions are formed in the P^+ and N^+ regions and there is a uniform electric field in the intrinsic layer. The electric field sweeps out the mobile carriers from the intrinsic layer and the layer is depleted of mobile carriers. When a reverse voltage is applied to the diode the space charge regions extend to very small distances in the heavily doped regions and the capacitance is practically constant. The reverse voltage increases the electric field intensity in the intrinsic layer. The effects of increase in field intensity such as ionisation is made use of in the IMPATT devices which we will be discussing in the next section.

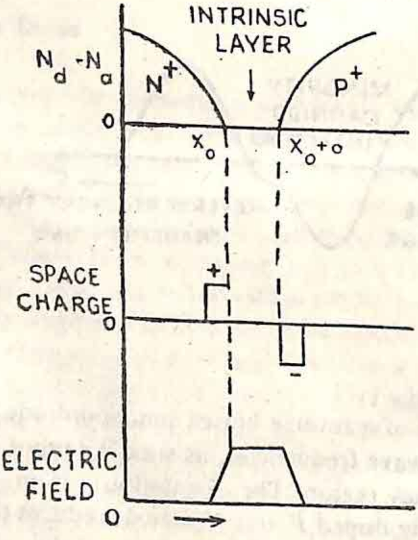


Fig. 6.7.

6.4.2 *NπP diode*

We consider an *NπP* diode for which the doping profile is given in Fig. 6.8.

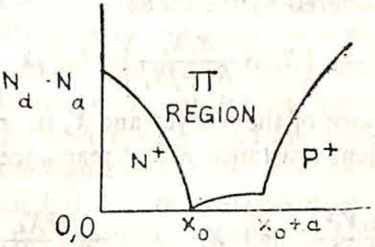


Fig. 6.8.

For this diode we have a high resistivity *P* layer between the heavily doped *N* and *P* regions. Since the π region is not intrinsic the depletion layer spreads into the region and the extend to which it spreads depends on the voltage across the diode. If *V* is the sum of the applied and barrier voltage and N_a the doping density in the π region the width of the depletion layer

$$x = \sqrt{\frac{2\epsilon V}{qN_a}} \tag{6.21}$$

The region is completely swept of mobile carriers for a voltage

$$V = \frac{qN_a \cdot a^2}{2\epsilon} \tag{6.22}$$

When the region is completely swept the diode has a minimum capacitance given by the expression

$$C = \frac{A\epsilon}{a} \tag{6.23}$$

a being the thickness of the π layer.

If the applied voltage is not sufficient to sweep the region of mobile carriers the region has a resistance and also capacitance. The resistance and capacitance combination appears in series with the junction.

If *W* is the width of the depletion layer, the resistance of the π region

$$R_i = \frac{\rho}{A}(a - W)$$

where ρ is the resistivity of the π region.

The junction capacitance $C_j = \frac{A\epsilon}{W}$

The capacitance of π region $C = \frac{A\epsilon}{a\ W}$

The impedance offered by the diode

$$Z = j\left(X_j + \frac{R_i X_i}{R_i + jX_i}\right) = R + jX \tag{6.24}$$

where X_i is the reactance of the π layer and X_j the reactance of the junction. The equivalent tesistance R and reactance X are given by the expressions

$$R = \frac{R_i X_i^2}{R_i^2 + X_i^2} \text{ and } X = X_j + \frac{R_i^2 X_i}{R_i^2 + X_i^2}$$

Substituting the values of R_i, X_i and X_j we find

$$R = \frac{\rho(a-W)}{A(\rho^2\omega^2\epsilon^2+1)} \tag{6.25}$$

$$X = \frac{-\rho^2\omega\epsilon}{A(\rho^2\omega^2\epsilon^2+1)}\left\{a + \frac{W}{\omega^2\epsilon^2\rho^2}\right\} \tag{6.26}$$

If the π region resistivity and the frequency of operation are high, so that $\rho^2\omega^2\epsilon^2 \gg 1$ we have

$$R \simeq 0 \tag{6.27}$$

$$X \approx \frac{-a}{\omega A\epsilon} \tag{6.28}$$

We thus see that the capacitance offered by the *PIN* diode at high frequencies is a constant depending only on the width of the intrinsic layer.

When The applied voltage is such that the π region is not swept out completely the diode offers a resistance which depends on the voltage as the width of the depletion layer W is a function of the applied voltage. The capacitance of the diode is a constant. Under these conditions the diode functions as a variolosser or attenuator.

The *PIN* diode when used as a switch operates between the ON and OFF states. In the ON state the diode is forward biased when minority carriers are injected into the high resistivity region between the highly doped regions. The increase in the densities of carriers reduces the resistance and the diode impedance is very low. In the OFF state the high resistivity region is completely swept out and the diode impedance is very high as the capacitance is very low.

6.4.3 *PIN diode operated at the breakdown voltage*

When the reverse voltage across a *PIN* diode is increased to such a value that the electric field intensity in the intrinsic region becomes sufficiently large to cause impact ionisation the diode offers negative resistance at microwave frequencies. The negative resistance of the diode implies a transfer of power from the biasing source to the microwave signal and the diode amplifies the signal when located in the transmission line. The amplification of the *PIN* diode when operated at the breakdown voltage is due to the impact ionisation within the depletion layer. It is physically not obvious but mathematical analysis brings out the negative resistance. The process of transfer of energy from the biasing source to the signal is obvious in a Read diode which we will be discussing in the next section.

It was shown that in a *PIN* diode the field intensity is constant within the intrinsic layer between the heavily doped *P* and *N* regions. When the reverse bias is large enough to cause avalanche breakdown electrons and holes injected into the intrinsic layer from the *N* and *P* sides travel with a velocity sufficient to cause impact ionisation. Due to avalanche multiplication the electron and hole currents reach large values. When a small alternating voltage is superimposed on the bias voltage the electric field intensity within the depletion layer is not a constant but varies about a mean value determined by the bias voltage. As the ionisation rate is a function of the field intensity the avalanche multiplication is modified and the diode offers negative impedance to the alternating signal.

To analyse the function of the diode in the breakdown region, consider a section *A* in the depletion layer of the diode.

If p and n are the densities of holes and electrons at the section and if it is assumed that the charge carriers move at the scattering limited velocity v_{sl} the current densities J_n and J_p at the section are given by the expressions

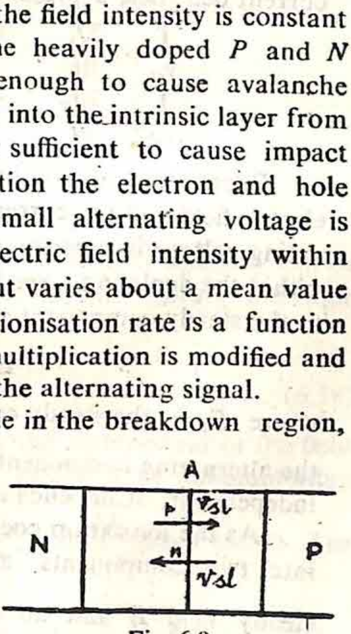


Fig. 6.9.

$$\left.\begin{aligned} J_n &= qv_{sl}\cdot n \\ J_p &= qv_{sl}\cdot p \\ J &= J_n+J_p = qv_{sl}\,(n+p) \end{aligned}\right\} \qquad (6.29)$$

Applying the Poisson's equation to the section at A we have

$$\frac{\partial E}{\partial x} = \frac{q(N_d - N_a + p - n)}{\epsilon}$$

$$= \frac{q(N_d - N_a)}{\epsilon} + \frac{1}{\epsilon v_{sl}}(J_p - J_n) \tag{6.30}$$

The continuity equations for holes and electrons at the section are

$$\left.\begin{aligned} \frac{\partial n}{\partial t} &= \frac{1}{q}\cdot\frac{\partial J_n}{\partial x} + \alpha v_{sl}(n+p) \\ \frac{\partial p}{\partial t} &= -\frac{1}{q}\,\frac{\partial J_p}{\partial x} + \alpha v_{sl}(n+p) \end{aligned}\right\} \tag{6.31}$$

where α is the ionisation coefficient. The recombination at the section is neglected as it is negligible compared to the multiplication due to ionisation. Expressing the carrier densities in terms of the current densities, expressions (6.31) are modified as

$$\left.\begin{aligned} \frac{1}{v_{sl}}\cdot\frac{\partial J_n}{\partial t} &= \frac{\partial J_n}{\partial x} + \alpha(J_n + J_p) \\ \frac{1}{v_{sl}}\cdot\frac{\partial J_p}{\partial t} &= \frac{-\partial J_p}{\partial x} + \alpha(J_n + J_p) \end{aligned}\right\} \tag{6.32}$$

Expressions (6.30) and (6.32) give the relation between the electric field and the current densities at the section. When an alternating voltage is superimposed on the bias voltage the electric field within the depletion layer has an alternating component in addition to the steady component and is given by the expression

$$E = E_0 + \widetilde{E}\exp j\omega t \tag{6.33}$$

where E_0 is the steady component and $\widetilde{E}$ is the maximum value of the alternating component at the section A. $\widetilde{E}$ is therefore time independent. It depends on the distance x.

As the ionisation coefficient α is field dependent it can be split into two components, a steady component α_0 determined by the steady field E and an alternating component $\frac{\partial \alpha}{\partial E}\cdot\widetilde{E}\exp j\omega t$. Therefore

$$\alpha = \alpha_0 + \frac{\partial \alpha}{\partial E}\cdot\widetilde{E}\exp j\omega t \tag{6.34}$$

The current densities J_n and J_p at the section contain steady and alternating components and can be written as

$$\left.\begin{aligned} J_n &= J_{n0}+\tilde{J}_n \exp j\omega t \\ J_p &= J_{p0}+\tilde{J}_p \exp j\omega t \end{aligned}\right\} \qquad (6.35)$$

where J_{n0} and J_{p0} are the steady values of the current densities and $\tilde{J}_n, \tilde{J}_p$ the maximum values of the alternating components. The alternating component of the current at the section is the sum of the drift component and the displacement current at the section. The displacement current is due to the alternating component of the electric field intensity and is given by the expression $j\omega\epsilon \tilde{E}$ exp $j\omega t$. The maximum value of the alternating current density at the section

$$\tilde{J} = \tilde{J}_n+\tilde{J}_p+j\omega\epsilon\tilde{E} \qquad (6.36)$$

The alternating component of the current density can be related to the field intensity by manipulating expressions (6.32), (6.34), (6.35) and (6.36). The relations are

$$\left.\begin{aligned} \frac{j\omega}{v_{sl}}\cdot \tilde{J}_n &= \alpha_0(\tilde{J}-j\omega\epsilon\tilde{E})+\frac{\partial\alpha}{\partial E}\cdot\tilde{E}\cdot J+\frac{\partial J_n}{\partial x} \\ \frac{j\omega}{v_{sl}}\cdot \tilde{J}_p &= \alpha_0(\tilde{J}-j\omega\epsilon\tilde{E})+\frac{\partial\alpha}{\partial E}\cdot\tilde{E}\cdot J-\frac{\partial J_p}{\partial x} \end{aligned}\right\} \qquad (6.37)$$

$J_{p0}+J_{n0}=J$ (steady current density).

The alternating component of the electric field intensity can be related to the alternating current densities by substituting the value of E from expression (6.33) in (6.30)

$$\frac{\partial\tilde{E}}{\partial x}=\frac{1}{\epsilon v_{sl}}(\tilde{J}_p-\tilde{J}_n) \qquad (6.38)$$

The relation between the alternating component of the field intensity and current density is now obtained by differentiating (6.38) and substituting the value of $\frac{\partial\tilde{J}_n}{\partial x}$ and $\frac{\partial\tilde{J}_p}{\partial x}$ from (6.37). The differential equation is

$$\frac{\partial^2\tilde{E}}{\partial x^2}+\frac{1}{\epsilon v_{sl}}\left(2\alpha_0 j\omega\epsilon-2\frac{\partial\alpha}{\partial E}J+\frac{\omega^2\epsilon}{v_{sl}}\right)\tilde{E}=\frac{\tilde{J}}{\epsilon v_{sl}}\left(2\alpha_0-\frac{j\omega}{v_{sl}}\right) \qquad (6.39)$$

The impedance offered by the diode is found by integrating the above expression to get the field intensity and calculating the alternating voltage across the diode from the value of the field intensity. The integration is difficult because of the factor in the brackets. However if it is assumed that the field within the depletion layer is constant the bracketed term is a constant and can be written as

$$k^2 = \left(2\alpha_0 j\omega\epsilon - 2\frac{\partial\alpha}{\partial E}J + \frac{\omega^2\epsilon}{v_{sl}}\right)\frac{1}{\epsilon v_{sl}} \tag{6.40}$$

On integrating the expression we get

$$\tilde{E} = C_1 \exp jkx + C_2 \exp(-jkx) + \frac{\dfrac{2\alpha_0}{\epsilon v_{sl}} - \dfrac{j\omega}{\epsilon v^2_{sl}}}{k^2}\tilde{J} \tag{6.41}$$

where C_1 and C_2 are the constants of integration to be obtained by numerical computation.

The alternating voltage is now obtained by integrating the above expression.

The current densities J_n and J_p can also be found from expressions (6.36) and (6.41)

$$\tilde{J}_n + \tilde{J}_p = \tilde{J} - j\omega\epsilon\tilde{E}$$

$$= -j\omega\epsilon\{C_1 \exp jkx + C_2 \exp(-jkx)\} - \frac{2\alpha'\tilde{J}J}{\epsilon v_{sl}k^2} \tag{6.42}$$

From expressions (6.38) and (6.41)

$$\tilde{J}_p - \tilde{J}_n = \epsilon v_{sl}\frac{\partial\tilde{E}}{\partial x}$$

$$= jk\epsilon v_{sl}C_1 \exp jkx - jk\epsilon v_{sl}C_2 \exp(-jkx) \tag{6.43}$$

From expressions (6.42) and (6.43) the hole and electron components of the alternating current are found as

$$\tilde{J}_n = \frac{j\epsilon v_{sl}}{2}\left(k - \frac{\omega}{v_{sl}}\right)C_1 \exp(jkx) - \frac{j\epsilon v_{sl}}{2}\left(k + \frac{\omega}{v_{sl}}\right)\exp(-jkx)$$
$$- \frac{\alpha' J_0}{k^2\epsilon v_{sl}}\cdot\tilde{J} \tag{6.44}$$

$$\tilde{J}_p = \frac{-j\epsilon v_{sl}}{2}\left(k + \frac{\omega}{v_{sl}}\right)C_1 \exp(jkx) + \frac{j\epsilon v_{sl}}{2}\left(k - \frac{\omega}{v_{sl}}\right)\exp(-jkx)$$
$$- \frac{\alpha' J_0}{k^2\epsilon v_{sl}}\cdot\tilde{J} \tag{6.45}$$

The small signal admittance of the diode is obtained as the ratio of the current to the voltage across its terminals

$$Y = \frac{A\tilde{J}}{\int_0^x \tilde{E}\,dx} \tag{6.46}$$

When the admittance is evaluated it is observed that the conductance part of the admittance is negative and this is the cause of energy transfer from the biasing field to the microwave signal. The susceptance part of the impedance changes from inductive to capacitive at a particular frequency referred to as the resonant frequency. The resonant frequency is in the Gigahertz range.

6.5.1 *Impact Avalanche transit time diode (IMPATT)—Read diode*

In the last section we discussed the operation of *PIN* diode near breakdown and found that diode offers negative resistance to microwave signals and as such can be used in amplifying circuits. However the physical mechanism of transfer of energy from the biasing source is not clear. In this section we consider the Read structure in which the transfer of energy becomes clear.

Before we take up for consideration the specific structure it is possible to state the conditions under which a transfer of energy from biasing source to a signal occurs. Consider a small region *AB* across which a steady voltage is applied. Due to the steady voltage a steady electric field is set up between *A* and *B*. If an alternating voltage is applied between the points *A* and *B* the field intensity changes from $E+\tilde{E}$ to $E-\tilde{E}$, where $\tilde{E}$ is the maximum value of the field due to the alternating voltage. If energy is to be transferred from D.C. source to the alternating field it is necessary to drive charge carriers from *A* to *B* when the alternating field opposes the movement of the charge carriers. For effective transfer of energy from the D.C. field to the signal two conditions are to be satisfied. They are: (a) there should be a large concentration of carriers at *A* at the beginning of the half cycle during which the A.C. field is opposite in direction to the steady field; (b) the carriers present at *A* should reach *B* before the end of the negative half cycle as otherwise power flows from the A.C. field. The second condition implies that the transit time of carriers through the drift region is less than half the period of the alternating voltage.

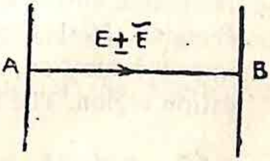


Fig. 6.10.

The above conditions are realised in a Read diode which has a doping profile as shown in Fig. 6.11.

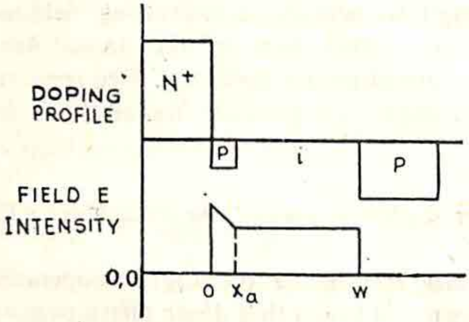


Fig. 6.11.

Due to the specific doping profile we have an intense electric field within the narrow *P*-region adjacent to the N^+ region. The depletion layer spreads throughout the narrow *P* region and the intrinsic region. Therefore the intrinsic region is depleted of mobile carriers and a constant field is set up within the intrinsic region. The field strength is sufficient to make the carriers move at the constant scattering limited velocity. The electric field intensity is maximum at the edge of the *P* region by the side of N^+ region and it decreases to the constant value within the narrow *P* region. In this region, the field intensity is high enough to cause impact ionisation, and the region is called the Avalanche multiplication region. The intrinsic region is called drift region.

6.5.2 *Avalanche region*

We have earlier calculated the time and space variations of hole and electron currents when there is avalanche multiplication in expressions (6.32). Adding the electron and hole currents we have

$$\frac{1}{v_{sl}}\cdot\frac{dJ}{dt}=\frac{d}{dx}(J_n-J_p)+2\alpha J \tag{6.47}$$

Integrating the above expression across the Avalanche region of thickness x_a we have

$$\frac{x_a}{v_{sl}}\cdot\frac{dJ}{dt}=\Big[J_n-J_p\Big]_0^{x_a}+2J\int_0^{x_a}\alpha\,dx \tag{6.48}$$

At $x = 0$ the hole component of the current J_p is due to holes moving into the Avalanche region from N^+ side and $J_p = J_{ps}$.

$$(J_n - J_p)_{x=0} = J - 2J_{ps}$$

At $x = x_a$ the electron component of the current is due to electrons moving into the Avalanche region from the right and $J_n = J_{ns}$

$$(J_n - J_p)_{x=x_a} = 2J_{ns} - J$$

Substituting the value of $(J_n - J_p)_0^{x_a}$ in expression (6.48) we have

$$\frac{x_a}{v_{sl}} \cdot \frac{dJ}{dt} = 2J\left(\int_0^{x_a} \alpha dx - 1\right) + 2\left(J_{ns} + J_{ps}\right)$$

Neglecting $(J_{ns} + J_{ps})$, the reverse saturation current in comparison with J and writing $\frac{x_a}{v_{sl}} = \tau$ (transit time through the avalanche region), we have

$$\tau \cdot \frac{dJ}{dt} = 2J\left(\int_0^{x_a} \alpha dx - 1\right) \tag{6.49}$$

The current density J in the avalanche region has two components, a steady component J_0 and an alternating component $\tilde{J} \exp(j\omega t)$, where $\tilde{J}$ is the maximum value of the alternating component. We can write

$$J = J_0 + \tilde{J} \exp j\omega t \tag{6.50}$$

The ionisation coefficient α is field dependent, however, as the region is very narrow we can associate an average value for the ionisation coefficient and write

$$\int_0^{x_a} \alpha dx = \bar{\alpha} x_a \tag{6.51}$$

As the field within the avalanche region is varying, the average value of the ionisation coefficient has two components, a steady value α_0 determined by the steady component of the electric field, and a varying component due to the alternating field. If the alternating component of the field has a maximum value $\widetilde{E}_a$, the average ionisation coefficient can be approximately written as

$$\bar{\alpha} = \alpha_0 + \alpha' \widetilde{E}_a \exp j\omega t \tag{6 52}$$

where $$\alpha' = \frac{\partial \alpha}{\partial E}$$

Substituting the value of α' in expression (6 51), we find

$$\int_0^{x_a} \alpha dx = \bar{\alpha} x_a = \alpha_0 x_a + \alpha' x_a \widetilde{E}_a \exp j\omega t \tag{6.53}$$

The steady field is of such a value as to cause impact ionisation, therefore, $\alpha_0 x_a = 1$.

Substituting the values of integral in expression (6.53) and the current density given in expression (6.50) in (6.49), we find

$$\tau j\omega \widetilde{J} = 2J\alpha' x_a \tag{6.54}$$

The maximum value of the current density due to holes and electrons at the edge of the depletion layer

$$\widetilde{J}_a = \frac{2J\alpha' x_a \widetilde{E}}{\tau \cdot j\omega} \tag{6 55}$$

$x_a \widetilde{E}$ is the alternating voltage across the avalanche region and the ratio of the alternating voltage to the particle current gives us the reactance of the avalanche region. If A is the sectional area of the diode, the reactance of the avalanche region

$$X_a = \frac{j\omega\tau}{2J\alpha' A}$$

The avalanche region has the properties of inductance so far as the particle current is concerned and we can associate an inductance L with the avalanche region

$$L_A = \frac{\tau}{2J\alpha' A} \tag{6.56}$$

In addition to the particle current there is a displacement current due to the varying field. The displacement current density

$$J_d = j\omega\epsilon \widetilde{E} \tag{6.57}$$

The ratio of the alternating voltage across the avalanche region to the displacement current gives us a reactance

$$X = \frac{\widetilde{E} x_a}{\omega\epsilon \widetilde{E} A} = \frac{x_a}{j\omega\epsilon A}$$

This reactance is capacitive and we can associate a capacitance in parallel to the inductance, the value of the capacitance being

$$C_A \equiv \frac{\epsilon A}{x_a} \tag{6.58}$$

The avalanche region thus functions as a parallel resonant circuit with a resonant frequency (ω_r)

$$\omega_r = \frac{1}{\sqrt{L_A C_A}} = \sqrt{\frac{2J\alpha' x_a}{\epsilon \tau}}$$

$$= \sqrt{\frac{2J\alpha' v_{st}}{\epsilon}} \tag{6.59}$$

The resonant frequency is normally in the Gigahertz range. The impedance of the avalanche region at a frequency ω can be readily found as

$$Z = \frac{1}{j\omega C_A} \cdot \frac{1}{\left(1 - \frac{\omega_r^2}{\omega^2}\right)} \tag{6.60}$$

The ratio of the particle current density J_A to the total current density

$$\gamma = \frac{J_A}{J_A + J_D} = \frac{1}{1 + \frac{J_D}{J_A}} = \frac{1}{1 - \omega^2 LC} = \frac{1}{\left(1 - \frac{\omega^2}{\omega_r^2}\right)} \tag{6.61}$$

At the resonant frequency ω_r the current through the avalanche region is zero, but at frequencies above the resonant frequency the current leads the voltage and we have a situation where the current has a large value at the instant the voltage across the avalanche region is less than its maximum value.

6.5.3 *Drift region*

The particles, holes and electrons move at the scattering limited velocity in the drift region, as such there is a continuous phase change with distance. If J_A is the particle current density at the edge of the avalanche region the particle current density at a distance x in the drift region is given by the expression

$$\tilde{J}_A(x) = \tilde{J}_A \exp \frac{j\omega(x - x_a)}{v_{st}} \tag{6 62}$$

The exponential terms represent the phase change. The displacement current density at the distance x

$$\tilde{J}_D(x) = \tilde{J} - \tilde{J}_A(x)$$
$$= \tilde{J}\left[1-\gamma \exp \frac{j\omega(x-x_a)}{v_{sl}}\right] \tag{6.63}$$

The displacement current density is related to the field intensity by the expression

$$\tilde{J}_D(x) = j\omega\epsilon\tilde{E}(x) \tag{6.64}$$

From expressions (6.63) and (6.64) we find the field intensity at a distance x in the drift region as

$$\tilde{E}(x) = \frac{\tilde{J}}{j\omega\epsilon}\left[1-\gamma \exp \frac{j\omega(x-x_a)}{v_{sl}}\right] \tag{6.65}$$

The voltage across the drift region V_d is now obtained by integrating the above expression between the limits x_a and W

$$\tilde{V}_d = \frac{\tilde{J}}{j\omega\epsilon}\left[(W-x_a)+\frac{\gamma v_{sl}}{j\omega}\left(\exp \frac{j\omega(W-x_a)}{v_{sl}}-1\right)\right] \tag{6.66}$$

$\frac{W-x_a}{v_{sl}}$ is the transit time of the carriers through the drift region τ_d, and $\omega\tau_d$ is the transit angle θ_d. The expression for voltage can also be written as

$$\tilde{V}_d = \frac{(W-x_a)\tilde{J}}{j\omega\epsilon}\left\{1-\frac{1}{\left(1-\frac{\omega^2}{\omega_r^2}\right)}\cdot\frac{1-\cos\theta_d}{j\theta_d}\right\} \tag{6.67}$$

The impedance of the drift region is now obtained as

$$Z = \frac{(W-x_a)}{1-\frac{\omega^2}{\omega_r^2}}\cdot\frac{1-\cos\theta_d}{\omega\epsilon A\theta_d}+\frac{W-x_a}{j\omega\epsilon A} \tag{6.68}$$

When the frequency of operation is above the resonant frequency of the avalanche region the current leads the voltage across the avalanche region and the current has a large value. The current pulse is driven through the drift region through a time less than half period when the transit angle $\theta_d<\pi$. Thus the conditions stipulated at the beginning of the section for transfer of energy from the biasing source to the signal are satisfied and the signal is amplified. Even though the theory of the Read diode is simple the diode is difficult to fabricate.

The Read diode and the *PIN* diode represent two extreme cases. In the *PIN* diode the avalanche multiplication occurs in the intrinsic layer and we do not have separate avalanche and drift regions whereas in the Read diode the avalanche multiplication is confined to a very narrow region.

A number of devices are fabricated which make use of the negative resistance due to interaction of the avalanche and drift regions. The devices are classified as Impact Avalanche Transit Time (IMPATT) diodes. The reliability and compactness makes them one of the powerful amplifying devices in the microwave region.

The statistical nature of avalanche multiplication is the cause of noise in IMPATT devices. The noise figure is found to depend on the variation of ionisation with field strength. It is experimentally found that IMPATT diodes using Ge and GaAs are quieter than the Si devices.

6.6.1 *Tunnel diode*

In this section we consider a diode which offers negative resistance under certain bias conditions. The negative resistance is due to tunneling and the diode is called a tunnel diode. It is used as an active device in electronic circuits in the frequency range of few hundred megahertz. Before we take up the principle of operation of the device we digress to discuss some properties of semiconductors with very high impurity concentration. The semiconductors are referred to as degenerate semiconductors.

It was proved in the first chapter that the Fermi level moves away from the mid band or intrinsic level when impurities are added. The extent to which the Fermi level moves is given by the expression

$$E_f \sim E_i = kT \ln \frac{N}{n_i} \tag{6.69}$$

where N is the impurity density and n_i the intrinsic carrier density.

When $E_f \sim E_i$ equals $E_g/2$ the Fermi level is moved to the band edge. For Ge and silicon the impurity densities necessary to move the Fermi level to the band edges can be found as $1.75 \cdot 10^{25}$ met^{-3} and $2.2 \cdot 10^{25}$ met^{-3}. If the impurity concentration exceeds the above values the Fermi level moves into the bands and the semiconductor is said to be degenerate The increase in the concentration of impurities to large values increases the densities of impurity states near the band edges and the impurity states also form bands. The bands due to the impurity states overlap with the energy bands causing what is known as band tailing. One of the consequences of band tailing is a reduction in the width of the forbidden gap.

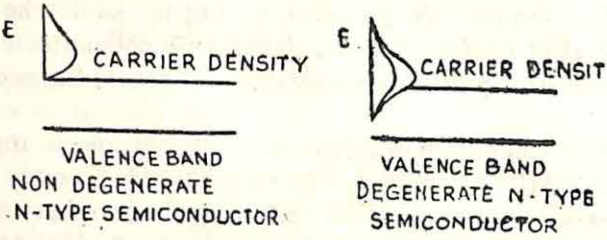


Fig. 6.12.

The band diagram of an isolated junction is given in Fig. 6.13. As the semiconductor regions on either side are degenerate the barrier voltage existing across the junction is more than the band gap (0.7 V for Ge and 1.10 for Si) and the depletion layer is very narrow (of the order of 50 Å).

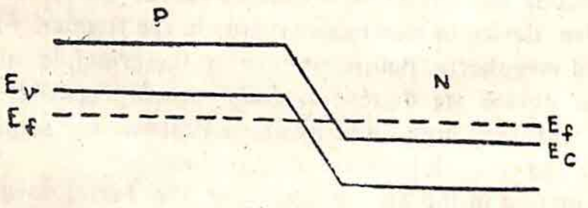


Fig. 6.13

From the band diagram we notice that the depletion layer separates electrons and vacant energy states having the same energy. The vacant energy states in the valence and conduction bands are separated by a potential barrier of height equal to the band gap and of width equal to the width of the depletion layer. Due to the wave nature of electrons they tunnel through the barrier from one side to another. The tunneling process results in carrier flow from the valence band on the *P* side to conduction band on the *N* side and vice versa (In a normal diode carriers move from the valence band to valence band and conduction band to conduction over the barrier.) The tunneling probability *P* is given by the expression

$$P=\exp\left(-\frac{4}{3}\sqrt{2m}\,\frac{E_g^{3/2}}{\hbar qE}\right) \tag{6.70}$$

The probability becomes appreciable when the field intensity reaches a value 10^8 V met^{-1} as is obtained in a junction of degenerate semiconductors.

The tunneling current depends on the product of the density of electrons on one side, the density of vacant energy states on the other side of the barrier and the tunneling probability. The density of electrons on the P side within the energy range E and $E+dE$ is $n_v(E)F_v(E)dE$, where $n_v(E)dE$ is the density of states within the energy range E and $(E+dE)$, and $F_v(E)$ is the probability of their occupation. The density of vacant energy states on the N side of the barrier is $n_c(E)\,[1-F_c(E)]\,dE$, where $n_c(E)\;dE$ is the density of energy states and $F_c(E)$ the probability of their occupation.

Therefore the tunneling current from P side to N side in the energy range E and $E+dE$

$$\propto Pn_v(E)F_v(E)n_c(E)[1-F_c(E)]\,dE \tag{6.71}$$

The total current due to tunneling from P side to N side is obtained by integrating the expression (6.71) between the limits E_c and E_v (Fig. 6.13)

$$I_{v\to c}\propto P\int_{E_c}^{E_v} n_v(E)F_v(E)[1-F_c(E)]n_c(E)\,dE \tag{6.72}$$

Similarly the tunneling current from N side to P side

$$I_{c\to v}\propto P\int_{E_c}^{E_v} n_c(E)F_c(E)[1-F_v(E)]n_v(E)\,dE \tag{6.73}$$

The total tunneling current through the junction

$$I=I_{c\to v}-I_{v\to c}\propto P\int_{E_c}^{E_v}[F_c(E)-F_v(E)]n_c(E)n_v(E)\,dE \tag{6.74}$$

In an isolated junction the Fermi level is invariant therefore $F_c(E)=F_v(E)$ and the total tunneling current is zero.

When a forward bias of V volts is applied the Fermi levels are separated by an energy gap qV and the band diagram is of the shape given in Fig. 6.14.

The tunneling current under a forward bias of V volts can be obtained from expression (6.74).

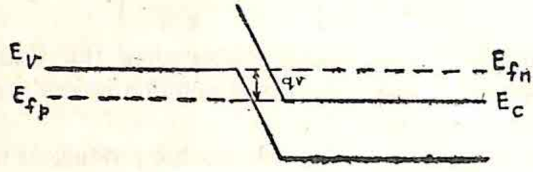


Fig. 6.14.

As the energy E is close to the Fermi level the Fermi function can be approximated as

$$F(E) \approx \frac{1}{2} - \frac{E-E_f}{kT} \tag{6.75}$$

$$F_c(E) - F_v(E) = \frac{E_{fn}-E_{fp}}{kT} \tag{6.76}$$

$E_{fn}-E_{fp}$ is the energy difference between the Fermi levels on the N and P sides and is equal to qV. Substituting the value of $F_c(E)-F_v(E)$ in expression (6.74) we find that the tunneling current is given by the expression

$$I = \text{const} \cdot \frac{qV}{kT} \int_{E_c}^{E_v} n_c(E) \cdot n_v(E)\, dE \tag{6.77}$$

$$n_c(E) \propto (E-E_c)^{1/2}$$

and

$$n_v(E) \propto (E_v-E)^{1/2}$$

where E_c and E_v are the energies corresponding to the band edges on the N and P sides. Integrating the expression (6.77) after substituting the values of $n_c(E)$ and $n_v(E)$ we find

$$I = \text{const} \cdot \frac{qV}{kT} (E_v-E_c)^2 \tag{6.78}$$

$$E_v-E_c = (E_v-E_{fp})+(E_{fn}-E_c)-(E_{fn}-E_{fp})$$

$$E_{fn}-E_{fp} = qV$$

$E_{fn}-E_c$ and E_v-E_{fp} represent the extents to which the semiconductors on either side of the junction are degenerated. They can be related to voltages V_n and V_p so that

$$E_v-E_{fp} = qV_p$$

and

$$E_{fn}-E_c = qV_n$$

On substitution of values (E_v-E_{fp}), $(E_{fn}-E_c)$ and $(E_{fn}-E_{fp})$ we find

$$(E_v-E_c)^2 = q(V_n+V_p-V)^2 \tag{6.79}$$

The tunneling current is given by the expression

$$I = \text{const} \cdot V(V_n+V_p-V)^2 \qquad (6.80)$$

The variation in current with applied voltage can be found by differentiating (6.80)

$$\frac{dI}{dV} = \text{const} \cdot \left[(V_n+V_p)^2+V\{3V-4(V_n+V_p)\}\right] \qquad (6.81)$$

The conductance is positive and the current increases with the applied voltage so long as $(V_n+V_p)^2 > V\{4(V_n+V_p)-3V\}$. The situation can be easily visualised if we remember that the densities of electrons and holes (vacant energy states) increase near the edges of the conduction and valence bands. For small forward voltages the current due to electrons tunneling from N side to P side is more than the current due to electrons tunneling from the P side to N side. The tunneling current reaches a maximum value when the applied forward voltage $V = \frac{V_n+V_p}{3}$. When the applied voltage increases beyond $\frac{V_n+V_p}{3}$ the energy states having the same energy as the electrons in the valence and conduction bands fall in the forbidden gap and the tunneling current decreases. The tunneling current becomes zero when the overlapping of the bands across the depletion layer disappears, i.e. when $V = V_n+V_p$. The diode offers negative resistance within the voltage range $\frac{V_n+V_p}{3}$ and V_n+V_p.

The theoretical and experimental characteristic curves of a tunnel diode are shown in Fig. 6.15.

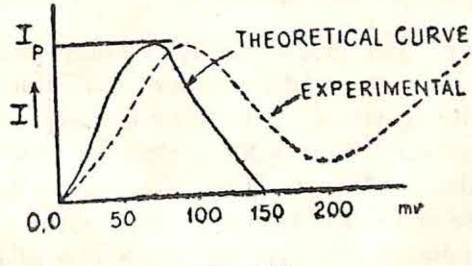


Fig. 6.15.

The current through the diode does not fall to zero value as indicated in the theory but reaches a minimum value The minimum value of current is due to: (a) band tail tunneling; and (b) exponential excess current by way of tunneling through energy states in the forbidden gap. Since tunneling is a quantum mechanical phenomenon there is no time lag between the application of voltage and consequent current variation as such the tunnel diode can be operated at very high frequencies.

6.6.2 *Equivalent circuit of tunnel diode*

The tunnel diode is operated in the negative resistance region as an amplifier or an oscillator and the equivalent circuit is given in Fig. 6.16.

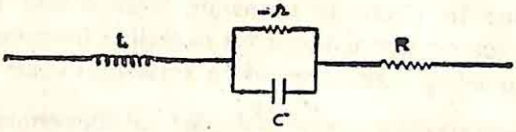


Fig. 6.16.

L is the lead inductance, and C is capacitance of the junction given by the expression

$$C = \sqrt{\frac{q\epsilon N^*}{2(V_b - V)}} \qquad (6.82)$$

$$N^* = \frac{N_a N_d}{N_a + N_d}$$

R_s is the bulk resistance of the semiconductor regions on either side of the junction.

6.7.1 *Ridley-Watkins-Hilsum (RWH) mechanism—Gunn effect*

Ridley-Watkins and independently Hilsum predicted that semiconductor materials under certain conditions can offer differential negative resistance. This differential negative resistance is a bulk effect and is due to transfer of electrons from one valley to another in the conduction band. The effect is experimentally observed by Gunn and is also known as Gunn effect.

The semiconductor material on which extensive studies are made is GaAs and we discuss the RWH mechanism and Gunn effect in GaAs. The $E-k$ diagram of GaAs (Fig. 6.17) has the following interesting characteristics due to which the Gunn effect is observed:

(a) In the conduction band there are two valleys separated by an energy gap of 0.36 eV.

(b) The effective mass of electrons in the lower valley is 0.072 m_0 while the effective mass of electrons in the upper valley is 1.2 m_0 (m_0 being the rest mass of electrons).

(c) The mobility of electrons in the lower valley is 500 cm^2/volt.sec while the mobility in the upper valley is 100 cm^2/volt.sec (the decrease in mobility is attributed to the larger effective mass).

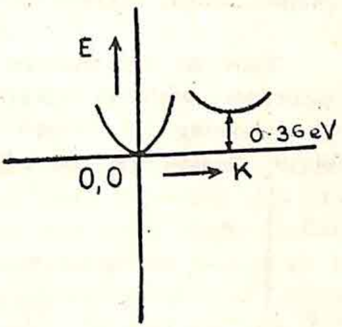


Fig. 6.17.

(d) The density of states in the upper valley is about 60 times the density of states in the lower valley.

Keeping in mind the band structure of GaAs we will study the current flow in the semiconductor. At room temperature all the electrons in the conduction band are located in the lower valley and the current density through the sample is given by the expression

$$J = q\mu_1 n_0 E \tag{6.83}$$

where μ_1 is the mobility in the lower valley and n_0 is the density of electrons. The velocity of electrons and with it the current density increases with the electric field intensity. This would have continued till the velocity of carriers reaches the scattering limited velocity had there been no upper valley situated close to the lower valley. Due to the presence of the upper valley the electrons are scattered into the upper valley after they reach a certain velocity. The movement of electrons into the upper valley slows them down and the mobility is reduced. The current is now determined by the mobilities in the two valleys and the densities of electrons in the valleys. If n_1 and n_2 are the densities of electrons in the lower and upper valleys and μ_1, μ_2 the mobilities in the two valleys the current densitv is given by the expression

$$J = q(\mu_1 n_1 + \mu_2 n_2)E \tag{6.84}$$

$$= qn_0\bar{\mu}$$

where $$\bar{\mu} = \frac{n_1\mu_1 + n_2\mu_2}{n_0}$$

ponential excess current by way of tunneling through energy states in the forbidden gap. Since tunneling is a quantum mechanical phenomenon there is no time lag between the application of voltage and consequent current variation as such the tunnel diode can be operated at very high frequencies.

6.6.2 *Equivalent circuit of tunnel diode*

The tunnel diode is operated in the negative resistance region as an amplifier or an oscillator and the equivalent circuit is given in Fig. 6.16.

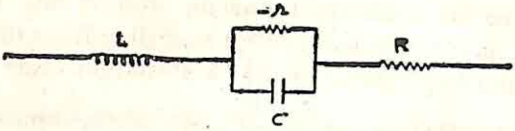


Fig. 6.16.

L is the lead inductance, and C is capacitance of the junction given by the expression

$$C = \sqrt{\frac{q\epsilon N^*}{2(V_b - V)}} \tag{6.82}$$

$$N^* = \frac{N_a N_d}{N_a + N_d}$$

R_s is the bulk resistance of the semiconductor regions on either side of the junction.

6.7.1 *Ridley-Watkins-Hilsum (RWH) mechanism—Gunn effect*

Ridley-Watkins and independently Hilsum predicted that semiconductor materials under certain conditions can offer differential negative resistance. This differential negative resistance is a bulk effect and is due to transfer of electrons from one valley to another in the conduction band. The effect is experimentally observed by Gunn and is also known as Gunn effect.

The semiconductor material on which extensive studies are made is GaAs and we discuss the RWH mechanism and Gunn effect in GaAs. The $E-k$ diagram of GaAs (Fig. 6.17) has the following interesting characteristics due to which the Gunn effect is observed:

(a) In the conduction band there are two valleys separated by an energy gap of 0.36 eV.

(b) The effective mass of electrons in the lower valley is 0.072 m_0 while the effective mass of electrons in the upper valley is 1.2 m_0 (m_0 being the rest mass of electrons).

(c) The mobility of electrons in the lower valley is 500 cm^2/volt.sec while the mobility in the upper valley is 100 cm^2/volt.sec (the decrease in mobility is attributed to the larger effective mass).

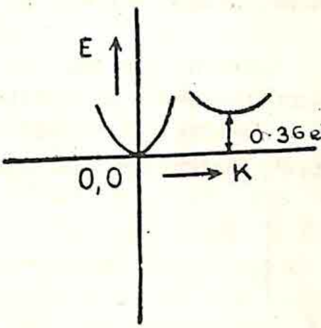


Fig. 6.17.

(d) The density of states in the upper valley is about 60 times the density of states in the lower valley.

Keeping in mind the band structure of GaAs we will study the current flow in the semiconductor. At room temperature all the electrons in the conduction band are located in the lower valley and the current density through the sample is given by the expression

$$J = q\mu_1 n_0 E \tag{6.83}$$

where μ_1 is the mobility in the lower valley and n_0 is the density of electrons. The velocity of electrons and with it the current density increases with the electric field intensity. This would have continued till the velocity of carriers reaches the scattering limited velocity had there been no upper valley situated close to the lower valley. Due to the presence of the upper valley the electrons are scattered into the upper valley after they reach a certain velocity. The movement of electrons into the upper valley slows them down and the mobility is reduced. The current is now determined by the mobilities in the two valleys and the densities of electrons in the valleys. If n_1 and n_2 are the densities of electrons in the lower and upper valleys and μ_1, μ_2 the mobilities in the two valleys the current densitv is given by the expression

$$J = q(\mu_1 n_1 + \mu_2 n_2)E \tag{6.84}$$

$$= q n_0 \bar{\mu}$$

where $$\bar{\mu} = \frac{n_1\mu_1 + n_2\mu_2}{n_0}$$

When all the electrons are transferred to the upper valley the current density is given by the expression

$$J = qn_0\mu_2 \tag{6.85}$$

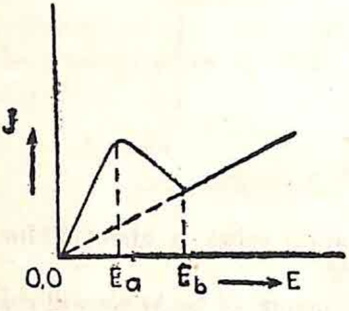


Fig. 6.18.

Thus we see that the current density varies with the field in accordance with the expressions (6.83), (6.84) and (6.85). From the plot showing the relation between current density and field intensity we find that the current density increases linearly till the field intensity reaches a value E_a. This is due to the fact that all electrons are confined to the lower valley till the field intensity reaches a value E_a. The current also increases linearly with the field intensity but with a reduced slope after the field intensity reaches a value E_b. E_b is the minimum value of the field intensity for which all the electrons are transferred to the upper valley and the current density is determined by the mobility in the upper valley. The semiconductor material exhibits differential negative resistance when the field intensity has a value intermediate to E_a and E_b. The current density through the material for a field intensity $E_a < E < E_b$ depends on the densities of electrons in the two valleys. Differentiating expression (6.84) with respect to E, we have

$$\frac{\partial J}{\partial E} = qn_0\frac{\partial \bar{\mu}}{\partial E} \tag{6.86}$$

$$\frac{\partial \bar{\mu}}{\partial E} = \frac{1}{n_0}\left(\mu_1\frac{\partial n_1}{\partial E} + \mu_2 \cdot \frac{\partial n_2}{\partial E}\right) \tag{6.87}$$

The differential $\frac{\partial \bar{\mu}}{\partial E}$ has two components $\mu_1\frac{\partial n_1}{\partial E}$ and $\mu_2\frac{\partial n_2}{\partial E}$.

As the field intensity increases the density of electrons in the lower valley decreases, and the density of electrons in the upper valley increases, therefore, $\partial n_1/\partial E$ is negative and $\partial n_2/\partial E$ is positive. Since $\mu_1 \gg \mu_2$ we find that the differential $\partial\bar{\mu}/\partial E$ has a negative value when the field intensity is between E_a and E_b. We need not go into the details of the transfer mechanism to study the principles of operation of devices based on electron transfer between valleys and consequential differential negative resistance. We assume that the mobility has an average negative value μ^* (about 100 cm^2/volt.sec) when the field intensity is between E_a and E_b.

6.7.2 *Dielectric relaxation time*

Any stray charge appearing at a point in a semiconductor decreases exponentially with a time constant called the dielectric relaxation time. We now evaluate the dielectric relaxation time for an N-type semiconductor of conductivity σ. If n is the density of charge carriers at a point in an N-type semiconductor the charge at the point decreases with time as electrons move away. If it is assumed that the diffusion is negligible compared to the drift we have from the continuity equation

$$\frac{dn}{dt} = \frac{1}{q}\frac{dJ_n}{dx}$$

$$= \mu_n n \cdot \frac{dE}{dx} \tag{6.88}$$

The charge density at the point is qn and we have from the Poisson's equation

$$\frac{dE}{dx} = \frac{-qn}{\epsilon} \tag{6.89}$$

Substituting the value of $\frac{dE}{dx}$ from expression (6.89) in (6.88)

$$\frac{dn}{dt} = \frac{-q\mu_n \cdot n \cdot n}{\epsilon}$$

$$= \frac{-\sigma \cdot n}{\epsilon} \tag{6.90}$$

Integrating the above expression with the limit that at $t = 0$, $n = n_0$ we have

$$n = n_0 \exp \frac{-t}{\tau} \tag{6.91}$$

Therefore $$Q(t) = Q_0 \exp \frac{-t}{\tau}$$

where $$\tau = \frac{\epsilon}{\sigma} \tag{6.92}$$

The dielectric relaxation time is very small, of the order of 10^{-12} secs, as such any localised charge disappears in a very short time and we are justified in saying that space charge neutrality is maintained in a semiconductor. If, however, the dielectric relaxation time is negative the localised stray charge instead of decreasing increases with time and gives rise to a series of pulses of current. This is Gunn effect in brief.

6.7.3 *Gunn effect*

Gunn while experimenting on a sample of *N*-type GaAs found that the current through the sample increased linearly with voltage till a certain threshold voltage. Beyond the threshold voltage a number of current pulses appeared with a time interval proportional to the length of the sample. The experimental facts can now be explained on the electron transfer mechanism and consequent differential negative resistance.

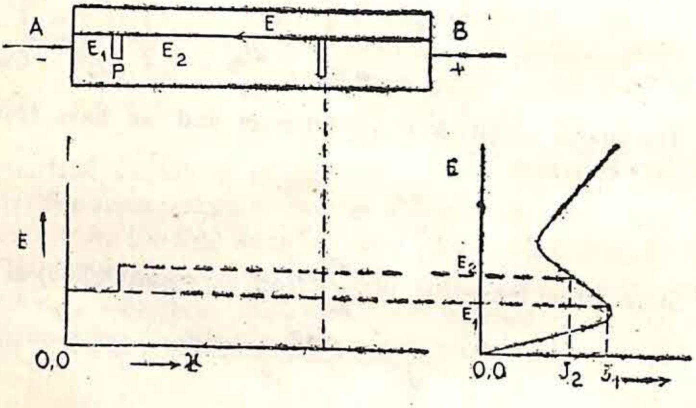


Fig. 6.19.

Consider a sample *AB* of *N*-type GaAs across which a voltage is applied. Under the influence of the electric field electrons drift across the sample and the drift current is proportional to the voltage. The dielectric relaxation time in the positive resistance region is positive and any charge nucleating in the bar dies out. When the electric field inside the bar is sufficient to give rise to negative differential resistance the dielectric relaxation time becomes negative and a charge nucleating in the bar grows instead of decaying while moving from *A* to *B*. The charge is referred to as an accumulation layer. The growth of the accumulation layer as it moves down the bar can be explained from the fact that the current density decreases instead of increasing in the negative resistance region. If P is an accumulation layer the electric field to the right of P is more than the field on the left. If the fields are projected on the current density field curve we notice that the current density on the left is more than the current density to the right; this causes an accumulation of electrons in the layer as it moves from A to B. If the sample is of sufficient length the accumulation layer stabilises at a value when the currents flowing into and out of the accumulation layer become equal. The layer moves at a velocity equal to the electron velocity in the sample and gives rise to a current pulse when it reaches B. The time interval between two pulses appearing at B is the time taken by the layer to traverse the bar and is equal to L/v_s, where L is the length of the bar.

If we have a small dipole layer nucleating at A the dipole grows as it drifts along the bar if the electric field in the bar is above the threshold value E_t. This growth of the domain can be explained as follows. As the field inside the domain is more than the field outside, we notice from the $E-J$ diagram that the current density inside the domain is less than the current outside. As the current densities are not equal, charges flow into the dipole layer and the dipole layer grows. The charge grows till the currents inside and outside become equal. When the domain stabilises the voltage across the sample is the sum of the voltages across the dipole and the remainder of the sample.

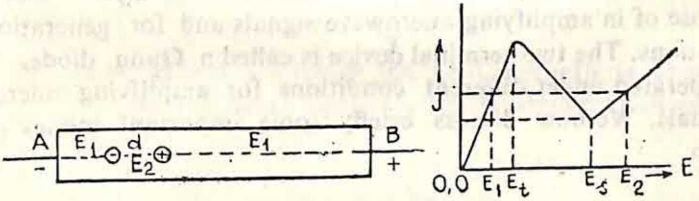


Fig. 6.20.

If E_2 is the average field within a dipole of length d and E_1 is the field outside the dipole, the voltage V across the sample is given by the expression

$$V = E_2 d + E_1(L-d)$$

From the above expression we find the dipole separation d as

$$d = \frac{V - E_1 L}{E_2 - E_1} \tag{6.93}$$

The dipole separation depends on the voltage across the sample. The growth or decay of the domain while shifting along the sample is determined by the voltage V. From the $E-J$ diagram it is obvious that the minimum field necessary to maintain a stable dipole is E_s. This field E_s is called the sustaining field and the voltage LE_s the sustaining voltage V_s.

The growth of an accumulation layer or a dipole is due to the differential negative resistance. The accumulation layer or dipole grows while moving along the sample as such the time during which they grow is the transit time within the sample L/v_s, where v_s is the drift velocity. We can now define a growth factor as the ratio of the size of the layer or dipole at the end of the bar B to its size at A. This growth factor is evidently exp $L/v_s\tau_d^*$, where τ_d^* is the dielectric relaxation time in the differential negative resistance region. We can now say that a stable domain or accumulation layer is formed when $L > v_s\tau_d^*$. Substituting the value of τ_d^* we have

$$Ln_0 > \frac{\epsilon v_s}{q\mu^*} \approx 10^{12} \text{ cm}^{-2} \tag{6.94}$$

The product of the sample length and the electron density is therefore an important factor determining the growth of domains or accumulation layer.

6.7.4 *Modes of operation*

The differential negative resistance offered by a sample of GaAs is made use of in amplifying microwave signals and for generation of oscillations. The two terminal device is called a Gunn diode. It can be operated under different conditions for amplifying microwave signals. We now discuss briefly some important modes of operation.

Limited space charge accumulation mode (LSA)

In the limited space charge accumulation mode of operation, the domain formation within the sample is inhibited and the field within the sample is not distorted by the domains. The device amplifies a microwave signal when it forms a part of a resonant cavity and biased such that the steady field in the sample is more than the threshold field as it offers differential negative resistance for field intensities more than the threshold value.

One of the methods of inhibiting domain formation is to keep the product of the electron density and sample length less than 10^{12} cm^{-2} when there cannot be any domain formation as discussed before. Even if the n_0L product is more than 10^{12} cm^{-2}, domain formation can be prevented by quenching the domains after their formation. Consider a sample across which a microwave signal of frequency f is applied. The electric field within the sample is the sum of the field due to the biasing voltage and the field due to the signal voltage. If the voltages are adjusted such that the field rises above the threshold value and falls below the sustaining field during a signal cycle, we find that during a part of the cycle conditions are favourable for domain formation and during the other part of the cycle domains are quenched.

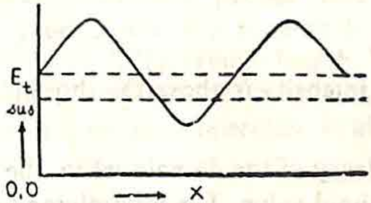


Fig. 6.21.

Domains grow when the dielectric relaxation time is negative and we can write down the growth factor as exp $(-\tau/\tau_d)$, where τ_d is the dielectric relaxation time. It has been shown earlier that $\tau_d = \epsilon/q\mu n$, where μ is the mobility and n the electron density. The mobility is field dependent, as such the growth factor G_1 can be written down as

$$G_1 = \exp \frac{-qn_0}{\epsilon} \int_0^{t_s} \mu dt \tag{6.95}$$

where t_s is the time during which the field strength is above the threshold value. The decay factor G_2 can be similarly written down as

$$G_2 = \exp\left(\frac{-qn_0}{\epsilon} \right) \int_{t_s}^{\tau} \mu dt \tag{6.96}$$

where $\tau = \dfrac{1}{f}$ the period of the signal cycle. The above integrals can be integrated only when the variation of mobility with field intensity is known. The integrals can be written as

$$G = \exp\left(-\frac{n_0}{f}\right) \cdot \frac{q}{\epsilon\tau} \int \mu dt \tag{6.97}$$

$\frac{q}{\epsilon}\int_{0}^{t_s} \mu dt$ represents the growth of the domain during the time the field intensity is above the threshold value while $\frac{q}{\epsilon}\int_{t_s}^{\tau} \mu dt$ represents the decay of the domain when the field intensity falls below the threshold value. The multiplying factor n_0/f determines the relative magnitudes of the growth and decay. Limits are set in the value n_0/f so that domain formation is inhibited and the field in the sample is not distorted by the domains. It is experimentally found that if n_0/f is within the limits 2×10^4 and 2×10^5 cm^{-3} sec, the sample can be used to amplify in the LSA mode. The mode diagram for the LSA mode is drawn as in Fig. 6.22. The efficiency of amplification in the LSA mode is about 20%.

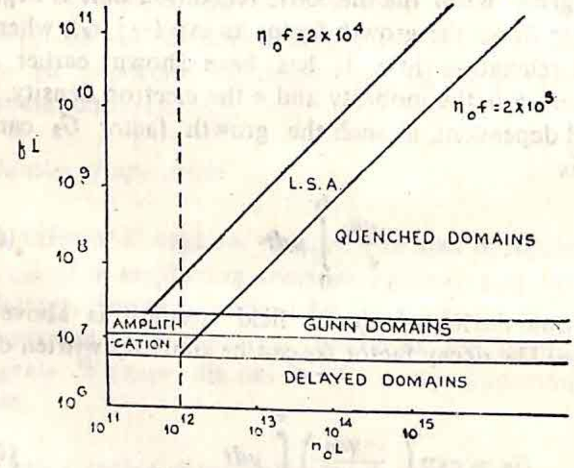


Fig. 6.22.

Transit time mode

If the frequency of the microwave signal is such as to make fL equal to the velocity with which domains travel in a sample (10^7 cm/sec) and if the field within the sample never falls below the sustaining voltage, domains nucleating at the cathode reach the anode once in every cycle and there is transfer of energy from the biasing field to the microwave signal. The efficiency of conversion will be maximum when the domains reach the anode during the negative half cycles of the signal. This can happen only when the domain width equals half the sample length. It is proved that the domain width depends on the n_0L product and this product is adjusted to obtain stable domains of half the sample length. This mode of operation is represented by a band about $fL = 10^7$ in the mode diagram. The efficiency in this mode of operation is about 10%.

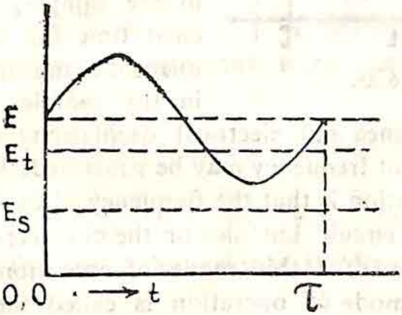


Fig. 6.23.

Delayed domains

If the period of the signal voltage is more than the transit time of the domain in the sample and further if the voltage falls below the threshold voltage a domain is formed in the sample only during the positive half of the signal cycle. During the formation of the domain and also during its transit through the sample the current through the device is maintained at a low value. The current suddenly rises to a large value and remains at the maximum value during the time the domain clears the anode. During this period the voltage across the sample is less than the threshold voltage, as such no domain is formed in the sample. The current wave form is a symmetrical square wave and the theoretical efficiency is about 27%. This mode of operation is referred to as delayed mode.

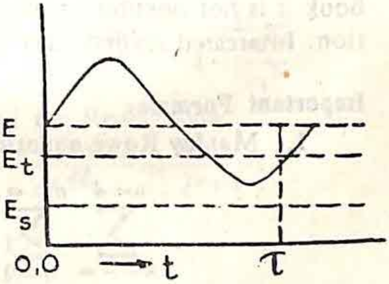


Fig. 6 24.

Quenched domain mode

If the frequency of the microwave signal is very high so that the signal period is much smaller than the transit time and if the bias voltage is adjusted such that the voltage across the sample falls below the sustaining voltage during a cycle, domains are formed and quenched in the sample. If there is not sufficient time for the domains to be quenched, multiple domains appear in the sample. The sample offers negative resistance and electrical oscillations of frequency many times the transit frequency may be generated. The advantage of this mode of operation is that the frequency of oscillation depends on the resonant circuit but not on the characteristics of the sample. The requirement for this mode of operation is for $fL > 2 \times 10^7$ cm/sec. This mode of operation is called the quenched domain mode.

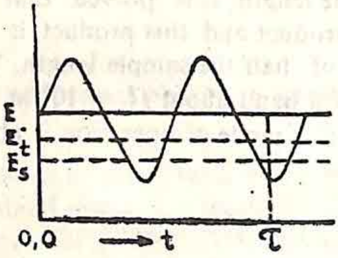


Fig. 6.25.

As is evident from the above discussion the Gunn diode is a versatile semiconductor device and can be used for amplification or for generation of microwave oscillations. In an introductory textbook it is not possible to discuss all the possible modes of operation. Interested readers may refer to advanced books.

Important Formulae

1. Manley Rowe equations

$$\sum_{n=-\infty}^{n=\infty} \sum_{m=0}^{m=\infty} \frac{mP_{mn}}{m\omega_1 + n\omega_2} = 0$$

$$\sum_{m=-\infty}^{m=\infty} \sum_{n=0}^{n=\infty} \frac{nP_{mn}}{m\omega_1 + n\omega_2} = 0$$

2. Capacitance of a junction $\propto V^{-1/(m+2)}$ for an arbitrary doping distribution

3. Cut off frequency of a varactor $\omega_{c0} = \frac{1}{C_j R_s}$

4. Dynamic cut off frequency of varactor

$$\omega_c = \frac{1}{R_s}\left(\frac{1}{C_{j\,\min}} - \frac{1}{C_{j\,\max}}\right)$$

5. Dynamic quality factor of a varactor $Q = \frac{1}{\omega R_s}$

6. Impedance of a $P\pi N$ diode $= R+jX$

where $R = \dfrac{\rho(a-W)}{A(\rho^2\omega^2\epsilon^2+1)}$

$$X = \frac{\rho^2\omega\epsilon}{A(\sigma^2\omega^2\epsilon^2+1)}\left\{a+\frac{W}{\omega^2\epsilon^2\rho^2}\right\} \approx -\frac{a}{A\omega\epsilon}$$

Particle current density equations when there is avalanche multiplication

$$\frac{1}{v_{sl}}\cdot\frac{\partial J_n}{\partial t} = \frac{\partial J_n}{\partial x} + \alpha(J_n+J_p)$$

$$\frac{1}{v_{sl}}\cdot\frac{\partial J_p}{\partial t} = -\frac{\partial J_p}{\partial x} + \alpha(J_n+J_p)$$

Inductance of the avalanche region in the Read diode

$$= \frac{\tau}{2J\alpha' A}$$

Capacitance $= \epsilon A/x_a$

Resonant frequency of the avalanche region $\sqrt{\dfrac{2J\alpha' x_a}{\epsilon\tau}}$

Ratio of particle current to total current $\gamma = \dfrac{1}{1-\dfrac{\omega^2}{\omega_r^2}}$

Impedance of the drift region of the Read diode

$$= \frac{W-x_a}{\left(1-\dfrac{\omega^2}{\omega_r^2}\right)}\cdot\frac{1-\cos\theta_d}{\omega\epsilon A\theta_d} + \frac{W-x_a}{j\omega\epsilon A}$$

Tunneling current $\alpha V(V_n+V_p-V)^2$

Dielectric relaxation time $\tau = \epsilon/\sigma$

Questions

1. What is meant by parametric amplification?
2. State Manley Rowe equations and explain on their basis: (a) up conversion, (b) amplification.
3. Derive an expression for the reverse biased junction capacitance when the doping is arbitrary. For which doping

profile does the capacitance vary inversely as the square of the applied voltage?

4. What are the figures of merit of varactor diode?
5. Prove that the dynamic quality factor of a varactor is the reciprocal of the product of the angular frequency and series resistance.
6. What is the important characteristic of a snap back diode? Give the typical doping profile for the diode.
7. Evaluate the impedance of a *PIN* diode.
8. Under what condition can a *PIN* diode be used as a switch?
9. Derive differential equations for particle densities when there is avalanche multiplication.
10. Derive the differential equations for the alternating field intensity in case of a *PIN* diode operated under breakdown conditions.
11. What are the conditions necessary for transfer of energy from a steady field to an alternating field? Explain how these are satisfied in a Read diode?
12. Show that the avalanche region of a Read diode has the characteristics of a parallel resonant circuit. Derive an expression for the resonant frequency.
13. Evaluate the impedance of the drift region in a Read diode.
14. Under what conditions will the drift region of the Read diode offer a negative resistance?
15. Explain why the Read diode offers positive resistance when the signal frequency is less than the resonant frequency of the avalanche region.
16. What is a degenerate semiconductor?
17. Derive an expression for the tunneling current in a tunnel diode.
18. Draw band diagrams and explain the phenomenon of tunneling.
19. Explain the origin of negative resistance in a bulk semiconductor.
20. What is meant by dielectric relaxation time? Evaluate it for an *N*-type semiconductor.
21. Explain the growth of: (a) accumulation layer, (b) a dipole in a bulk GaAs semiconductor.
22. State the condition for stable domain formation.
23. Discuss briefly the modes of operation of a Gunn diode.

7. Opto Electronic Devices

In this chapter we study some semiconductor opto electronic devices. Opto electronic devices are extensively used as light detectors and as sources of light. When used as a photodetector the device converts light energy into electrical energy—one of the extremely useful photodetector is the solar cell which is used to generate electric power. Among the light sources we have the semiconductor laser which converts electrical energy into coherent radiation. The existence of an energy gap in a semiconductor is the basic property made use of in the opto electronic devices. We first study the devices based on absorption of light and subsequently take up devices used in converting electrical energy to light energy.

7.1.1 *Photoconductivity*

We know that electrons are excited across the band gap in an intrinsic semiconductor due to the energy they acquire from thermal vibrations and the creation of electron hole pairs in a semiconductor is called thermal generation. Electrons are also excited across the band gap leaving holes in the valence band when they receive the requisite energy from light incident on the material. This type of generation is called optical generation. For optical generation of charge carriers, it is necessary that energy of the photon $h\nu$ is more than the forbidden gap. If the frequency of the incident light is such that $h\nu$ is less than the band gap the light is not absorbed and the material is transparent. It is interesting to note that all transparent substances are either insulators or wide band semiconductors.

One of the methods of determining the band gap in a semiconductor is to plot the absorption coefficient of the material as a function of the frequency. From the plot we find that the absorption coefficient suddenly rises at a frequency ν when $h\nu$ becomes equal to E_g.

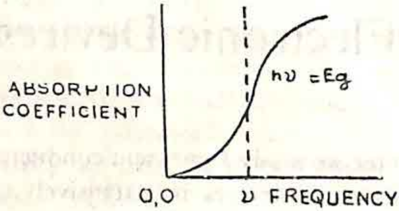


Fig. 7.1.

If φ is the number of photons incident on unit area of the material in unit time (units photons/cm^2. sec) the number of hole electron pairs generated in one second or the generation rate

$$g = \eta\varphi \tag{7.1}$$

where η is a constant called quantum efficiency. Alongside carrier generation there is recombination in the semiconductor and equilibrium is reached when the generation rate equals the recombination rate. If δn and δp are the excess densities of electrons and holes in a sample exposed to light, we have

$$g = \frac{\delta n}{\tau_n} = \frac{\delta p}{\tau_p} \tag{7.2}$$

where τ_n and τ_p are the lifetimes of electrons and holes in the sample.

The increase in conductivity of the sample due to the increase in carrier densities is given by the expression

$$d\sigma = qg(\tau_p\mu_p + \tau_n\mu_n) \tag{7.3}$$

7.1.2 *Photoconductors*

Photoconductors are semiconducting devices used for light detection. They make use of the property of photoconductivity and are also referred to as light detecting resistors (LDR). They are constructed by depositing the semiconducting material on a ceramic substrate. Contacts are made at the ends as shown in Fig. 7.2.

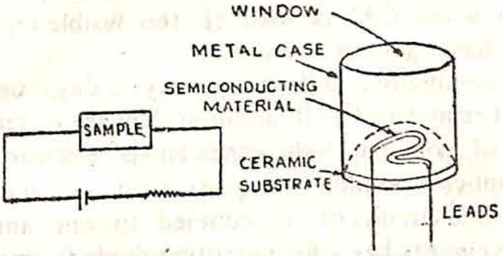


Fig. 7.2.

If the sample length is L, surface area A the number of hole electron pairs generated per unit thickness of the sample is LAg, where g is the generation rate. From expression (7.2) we find that the number of holes and electrons in the sample increases by an amount $LAg\tau$. If one of the carriers only contributes to the current as is often the case (one of the carriers is trapped) the excess charge due to the excess density is $qLAg\tau$. The increase in current through the sample dI can now be obtained as the ratio of the excess charge to the transit time t

$$dI = \frac{qLAg\tau}{t} \tag{7.4}$$

The gain of the photoconductor is defined as the ratio of the increase in current to the total charge generated by optical excitation $qLAg$ and is given by the expression

$$G = \frac{\tau}{t} \tag{7.5}$$

The transit time t is the ratio of the length of the sample to the velocity of charge carriers. $v = \mu E = \mu V/L$, where V is the voltage across the sample. Substituting the value of the transit time in expression (7.5) we find that

$$G = \frac{\tau \cdot \mu \cdot V}{L^2} \tag{7.6}$$

From the above expression we notice that the gain is inversely proportional to the square of the length. If L is decreased to increase the gain the surface area of the sample decreases and there is a decrease in sensitivity. The gain is increased by increasing the lifetime and selecting materials in which the carriers have a high mobility. One of the considerations in selecting the semiconductor material is its spectral response. Silicon, germanium and GaAs are used in the extreme red region, PbS and InSb are used in the infrared region while CdS is used in the visible region. The photoconductor has a gain of about 10^5.

Photoconductors find many day-to-day uses in exposure meters and relay circuits. In addition they are increasingly used for detection of coherent light waves and in electronic circuits incorporating optical coupling.

Electronic circuits can be coupled to one another optically. One of the circuits has a light emitting diode from which an intensity modulated light wave emanates. The light waves from the source fall on a photo detector and the signal is retrieved.

The photon flux density in the emitted light can be expressed as

$$\varphi = \varphi_0 + \varphi_1 \exp j\omega t \tag{7.7}$$

When the light falls on a photoconductor the current in the conductor has a steady component and an alternating component. The steady component of the photon density generates a charge $\eta qA\varphi_0\tau$ in the photoconductor and current due to the photon density is given by the expression

$$I_0 = \frac{\eta qA\varphi_0\tau}{t} \tag{7.8}$$

where t is the transit time through the detector. The alternating component of the photon density generates carriers in the sample proportional to the light intensity and the number of carriers generated is time dependent and can be expressed as $\delta p_\omega \exp j\omega t$. From the continuity equation we find

$$j\omega\delta p_\omega \exp j\omega t = \frac{-\delta p_\omega}{\tau} \exp j\omega t + \eta qA\varphi_1 \exp j\omega t$$

From the above expression we find

$$\delta p_\omega = \frac{\eta A\varphi_1\tau}{(1+j\omega\tau)} \tag{7.9}$$

The alternating component of the current is now obtained as

$$I_\omega = \eta qA\varphi_1 \cdot \frac{\tau}{t} \cdot \frac{1}{(1+j\omega\tau)} \tag{7.10}$$

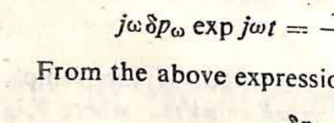


Fig. 7.3.

The equivalent circuit of the photoconductor when used in optically coupled circuit is as shown in Fig. 7.3. G in the equivalent circuit is the r.f. conductance of the sample. The average power output to a matched load is obtained as

$$P_{av} = \frac{1}{8}(I_\omega)^2 \cdot \frac{1}{G}$$

$$\approx \frac{1}{8}(\eta\, q\, \varphi_1 A)^2 \cdot \frac{1}{G(\omega t)^2}\bigg|_{\omega t \gg 1} \qquad (7.11)$$

We will see later that the signal power output from a photoconductor is much less than the power obtained from a photodiode.

The important considerations in a photoconductor are: (a) dark resistance; and (b) response time. The dark resistance is the resistance of the sample in the absence of illumination. It can be increased by increasing the length of the sample, which is not advisable as it adversely affects the gain, some compromise is made in deciding the length of the sample. The response time depends on the lifetime of the carriers in the sample. We again see that when the lifetime is increased the response time increases as,such one has to sacrifice gain for the speed of response.

7.2.1 *Photodiodes*

Photodiodes convert light energy incident on them to electrical energy and the effect is called photo voltaic effect. They are used as sources of power (solar cell) and also as photodetectors. Before we proceed to discuss the specific use we consider some properties of an illuminated junction.

When discussing the behaviour of a junction we noted that a barrier voltage exists across the depletion layer of a *PN* junction and the voltage opposes the diffusion of majority carriers from one side to another. The barrier voltage is of such magnitude and direction as to maintain the minority carrier densities in the bulk regions at the thermal equilibrium values. The barrier voltage cannot be measured as the contact potentials are opposite in direction and the sum of the contact potentials and barrier voltage is zero.

When a junction is illuminated the minority and majority carrier densities increase in the bulk semiconductor on either side of the depletion layer due to optical generation. As the field within the junction is from the *N* side to *P* side the holes generated on the *N* side move towards the *P* side where they become majority carriers. Similarly electrons generated on the *P* side move towards *N* side and become majority carriers.

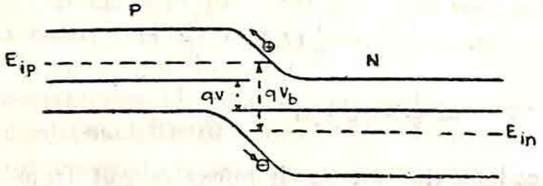


Fig. 7.4.

When the junction is open there cannot be any current, as such there is no movement of optically generated carriers from one side to another. The flow of optically generated carriers is opposed by an additional voltage appearing across the junction. The additional voltage is opposite in direction to the barrier voltage as it has to oppose the transfer of minority carriers from one side to another. In effect the barrier voltage is reduced due to optical generation and the change in barrier voltage appears as a measurable voltage across the junction. The open circuit voltage can never be greater than the energy gap.

When the junction is shorted the minority carrier densities at the edges of the depletion layer and the barrier voltage fall to thermal equilibrium values. The minority carrier densities in the bulk regions are more than the thermal equilibrium values due to optical generation and the fall in minority carrier densities to thermal equilibrium values at the edges of the depletion layer gives rise to a gradient and a current flows out of the junction. This current is due to optical generation within the bulk semiconductor region.

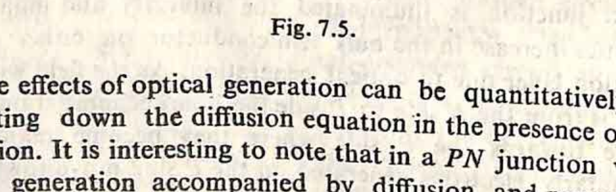


Fig. 7.5.

The effects of optical generation can be quantitatively found by writing down the diffusion equation in the presence of optical generation. It is interesting to note that in a *PN* junction we have optical generation accompanied by diffusion and recombination while in photoconductor the optical generation is accompanied by recombination and drift. If g is the generation rate the diffusion equation for holes on the N side is

$$D_p \frac{d^2(\delta p)}{dx^2} = -\frac{\delta p}{\tau_p} + g \tag{7.12}$$

Solving the equation for δp we find that

$$\delta p = C \exp \frac{-x}{L_p} + g\tau_p \tag{7.13}$$

The constant C in equation (7.13) is to be evaluated from the boundary conditions prevailing at the junction. It was shown in Chapter 2 that the excess carrier densities at the edge of the depletion layer ($x=0$) is related to the voltage applied to the junction, v by the expression

$$p_n(O) = p_{no}\left(\exp \frac{qV}{kT} - 1\right) \tag{2.33}$$

From the above boundary condition C can be found as

$$C = p_{no}\left(\exp \frac{qV}{kT} - 1\right) - g\tau_p \tag{7.14}$$

and

$$\delta p_n(x) = \left[p_{no}\left(\exp \frac{qV}{kT} - 1\right) - g\tau_p\right] \exp \frac{-x}{L_p} + g\tau_p \tag{7.15}$$

Similarly for electrons on the P side we have

$$\delta n_p(x) = \left[n_{po}\left(\exp \frac{qV}{kT} - 1\right) - g\tau_n\right] \exp \frac{x}{L_n} + g\tau_n$$

When the applied voltage is zero, i.e. when the diode is shorted the excess hole and electron densities on the N and P sides of the junction are zero. The densities of minority carriers in the bulk regions are more than the thermal equilibrium values due to optical generation. The carrier density profiles are given by the expression (7.15).

The current flowing when the junction is shorted can be obtained from expression (7.15) by setting $V=0$, we have

$$\left.\begin{aligned} I_p &= -qAD_p \left.\frac{dp}{dx}\right|_{x=0} = -qAgL_p \\ I_n &= -qAD_n \left.\frac{dn}{dx}\right|_{x=0} = -qAgL_n \end{aligned}\right\} \tag{7.16}$$

The total current $I_L = I_p + I_n$

$$= qAg\,(L_p + L_n) \tag{7.17}$$

flows out from the P side of the junction.

From the above expression we find that the current is due to carriers generated within a diffusion length on either side of the depletion layer. The carriers generated in the bulk semiconductor at distances more than the diffusion length do not contribute to the current as they are lost in recombination. The current voltage relation for an illuminated junction can be obtained from expression (7.15) as

$$I = qA\left(\frac{D_p p_{no}}{L_p} + \frac{D_n n_{po}}{L_n}\right)\left(\exp\frac{qV}{kT} - 1\right) - I_L \qquad (7.18)$$

It has been proved in Chapter 1 that

$$qA\left(\frac{D_p p_{no}}{L_p} + \frac{D_n n_{po}}{L_n}\right) = I_0 \text{ (Reverse saturation current)}$$

Substituting the value of I_0 in expression (7.18) we have

$$I = I_0\left(\exp\frac{qV}{kT} - 1\right) - I_L \qquad (7.19)$$

The current through the junction has two components—a component I_L due to optical generation flowing from N side to P side within the junction and a component $I_0\left(\exp\frac{qV}{kT} - 1\right)$ due to the forward bias appearing on the junction flowing from P side to N side.

The above expression enables us to find the open circuit voltage as

$$V_{oc} = \frac{kT}{q}\ln\left(\frac{I_L}{I_0} + 1\right) \qquad (7.20)$$

The I-V characteristic of the illuminated junction can be drawn from the expression (7.19).

The characteristic passes through the fourth quadrant indicating that power is delivered by the diode.

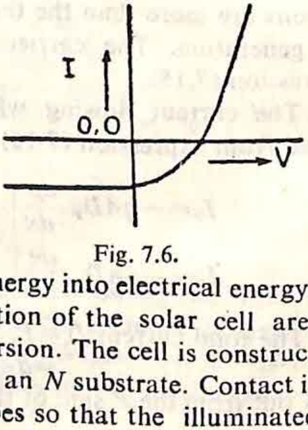


Fig. 7.6.

7.2.2 *Solar cell*

The ability of an illuminated junction to deliver power is made use of in a solar cell, used to convert solar energy into electrical energy. The considerations involved in the operation of the solar cell are: power delivered and efficiency of conversion. The cell is constructed by depositing a thin P layer on an N substrate. Contact is made to the P layer through narrow stripes so that the illuminated

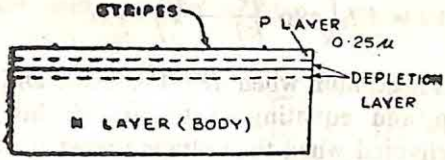


Fig. 7.7.

area is not reduced by the presence of contacts. The *P* layer thickness should be less than a diffusion length because the incident radiation should reach the junction and because the photo generated current I_L is due to carriers generated within a diffusion length on either side of the depletion layer.

The solar radiation has a wide frequency spectrum extending from infrared to ultraviolet. As the function of the cell is to convert maximum solar energy it suggests that the semiconductor used should have a narrow band gap However, when a narrow band gap semiconductor is used the high frequency radiation is strongly absorbed at the surface and the photo-generated carriers do not reach the junction as they are absorbed near the surface. The open circuit voltage is also less for a narrow band gap semiconductor. In effect the solar energy converted is restricted to a narrow range of wavelengths. The wide band semiconductor GaAs is used for solar cells operated in outer space while Si is used for solar cells on the surface of the earth. One other advantage Si offers over is GaAs that the diffusion lengths for holes and electrons are larger. Extensive search is being made for materials that can absorb maximum solar energy.

It was stated in the last section that an illuminated junction is capable of delivering power when operated in the fourth quadrant, i.e. when the voltage across the diode is less than the open circuit voltage. An ideal illuminated junction can be represented by an equivalent circuit as under.

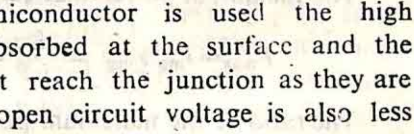


Fig. 7.8.

There are two current generators in the equivalent circuit—a generator representing the photo-generated current and another representing the forward biased current. If *V* is the voltage across the junction the power delivered, *P* is given by the expression:

$$P = VI = VI_0\left(\exp\frac{qV}{kT} - 1\right) - VI_L \tag{7.21}$$

The power is maximum when dP/dV is zero. Differentiating the above expression and equating it to zero we find that the maximum power is delivered when the voltage across the diode is V_{mp}. The voltage at which the power delivered is maximum is obtained from the expression

$$\left(1+\frac{q}{kT}V_{\text{mp}}\right)\exp\frac{qV_{\text{mp}}}{kT} = 1+\frac{I_L}{I_0} \tag{7.22}$$

The current when the power output is maximum I_{mp} is obtained from equations (7.21) and (7.22) as

$$I_{\text{mp}} = \frac{I_0 q}{kT}\cdot V_{\text{mp}}\cdot\exp\frac{qV_{\text{mp}}}{kT}. \tag{7.23}$$

The maximum power that can be obtained from the solar cell

$$P_{\text{max}} = I_{\text{mp}}V_{\text{mp}} = \frac{I_0 q}{kT}\cdot V^2_{\text{mp}}\cdot\exp\frac{qV_{\text{mp}}}{kT} \tag{7.24}$$

The ratio of the maximum power output to the incident optical power input gives us the efficiency of power conversion. If P_{in} is the optical power incident on the diode in watts/sq cm the optical power input is $P_{\text{in}}A$ and the efficiency

$$\eta = \frac{I_0 q}{kT}\;\frac{V^2_{\text{mp}}}{P_{\text{in}}A}\cdot\exp\frac{qV_{\text{mp}}}{kT}$$

$$= \frac{I_0 q}{AkT}\;\frac{I_L V^2_{\text{mp}}}{\left(1+\frac{q}{kT}V_{\text{mp}}\right)}\left(1+\frac{I_0}{I_L}\right)\frac{1}{P_{\text{in}}} \tag{7.25}$$

The solar power density is about 100 mw/sq cm and the theoretical efficiency is about 22 to 28% for Si and GaAs cells. The practical efficiency is about half the theoretical value due to various losses. One of the major losses is the reflection at the surface of the cell. Anti-reflection coatings are given to minimise the loss. One other loss is the series resistance which we will discuss in some detail.

The cell always has a resistance due to thickness of the semiconductor material on either side of the junction. As one of the regions is made thin for the incident radiation to reach the junction the series resistance is determined by the thin layer. If R_s is the series resistance of the cell, the equivalent circuit has the form as in Fig. 7.9.

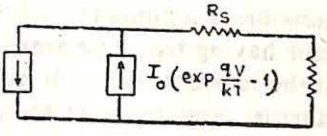


Fig. 7.9.

The current I in the circuit is now given by the expression

$$I = I_0\left[\exp\frac{qV-IR_s}{kT}-1\right]-I_L \qquad (7.26)$$

The power output is considerably reduced due to the series resistance. The ratio of the output power to the maximum available power is about 0.7 for a series resistance of 1 ohm while the ratio becomes 0.1 for a resistance of 10 ohm. As the series resistance arises out of the bulk resistance of the thin P layer on the top a compromise has to be made between optical absorption which increases as the layer thickness is increased and the series resistance which decreases with the increase in thickness.

7.2.3 *Photodiode detectors*

Diodes operated as detectors of radiation are all reverse biased. When radiation is incident on the diode the photo-generated carriers in the depletion layer and in the bulk semiconductors on either side of the depletion layer move under the influence of the electric field present in the depletion layer. The increase in reverse current is proportional to the intensity of the incident radiation.

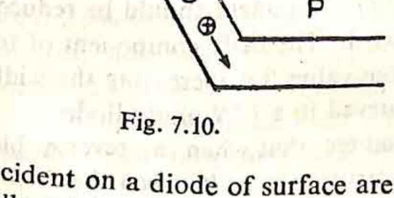


Fig. 7.10.

The carriers move in the depletion layer at high velocities and there is not much time for recombination. We can say that all the photo-generated carriers are collected by the reverse biased junction.

If a photon flux φ is incident on a diode of surface area A the number of charge carriers collected by the junction is equal to the number generated and the increase in the reverse current is $q\eta A\varphi$ (η being the quantum efficiency). When we compare the increase in current due to photogeneration in a diode to the increase in current in a photoconductor having the same area and same quantum efficiency we find that the current in the diode is t/τ times larger in the diode. This factor is very large as the transit time t is made larger than the lifetime of the carriers. The photodiodes are more sensitive than the photoconductors.

When operated as detectors of intensity modulated signals we have to take into consideration the capacitance of the reverse biased junction and the equivalent circuit of the diode when used to detect intensity modulated signals is given in the Fig. 7.11 (R is the series resistance of the diode and C the junction capacitance).

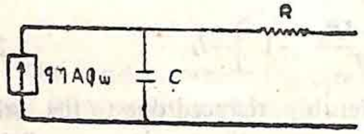


Fig. 7.11.

The maximum signal power available from the diode can easily be obtained from the equivalent circuit as

$$P = |\, q\eta A\varphi(\omega)\,|^2 \cdot \frac{1}{\omega^2 C^2 R} \tag{7.27}$$

Comparing the power output from a photoconductor (7.11) we find that the output from the diode is many times larger than that from the photoconductor.

The photo-generated current in a diode has two components—a diffusion component and a drift component. The diffusion component of the current is due to the diffusion of photo-generated carriers within the bulk semiconductor regions on either side of the depletion layer, while the drift current is due to the carriers generated within the depletion layer. If the diode is to be used for detecting intensity modulated signals the speed of response should be high, as such the diffusion component should be reduced and the drift component increased. The drift component of the current can be increased to a large value by increasing the width of the depletion layer. This is achieved in a PIN photodiode.

It was seen in the last chapter that when a reverse bias is applied to a PIN diode the intrinsic layer between the P^+ and

N^+ regions is depleted and we have a depletion layer of width equal to the thickness of the *i* layer. One other advantage of the *PIN* diode is that the width of the depletion layer and with it the capacitance can be adjusted to the desired value.

The *PIN* photodiode has a thin *N* layer formed on the intrinsic layer above a *P* substrate. External contact is made to the *N* layer by stripes on the surface. The light incident on the surface reaches the depletion layer through the *N* layer and electron hole pairs are generated within the depletion layer. It can easily be shown that the drift current is very large compared to the diffusion current.

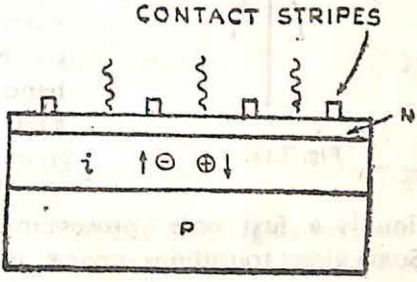


Fig. 7.12.

While evaluating the response to intensity-modulated signals one has to take into consideration the transit time of the carrier and also the phase change introduced due to definite transit time, in addition to the capacitance. The thickness of the intrinsic layer is tailored for specific applications.

Certain amount of amplification of the light signal is possible by increasing the reverse voltage to a value sufficient to cause avalanche multiplication, the resulting diode is called avalanche photodiode.

7.3.1 *Luminescence*

The light emission by materials is referred to as luminescence and is due to electron transitions from a higher energy state to a lower energy state when energy is released in the form of light. There are several types of luminescence depending on the method of excitation of electrons to higher energy states. In this section we discuss the luminescence caused by radiative recombination of electron hole pairs. This luminescence is referred to electroluminescence as the excitation is caused by current.

Radiative recombination is a direct transition in the sense the electron just falls into an available vacant energy state releasing

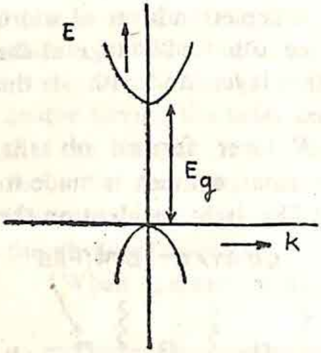


Fig. 7.13.

energy in the form of light. As momentum is to be conserved in any transition radiative recombination can occur only when the momenta of the electron in the conduction band and the vacant energy state are equal. In direct band gap semiconductors like GaAs the energy minimum in the conduction band and the energy maximum in the valence band occur at the momentum value $k=0$ as indicated by the $E-k$ diagram.

Therefore radiative recombination is a first order process in direct band gap semiconductors. Some direct transitions occur in other semiconductors between the conduction band and energy states having the same momentum called luminous centres.

In a nonradiative recombination process the electron transition to a lower energy state involves phonons and as such the energy released is converted to heat.

We now consider the recombination process to evaluate the ratio of the radiative recombination rate to the generation rate. The ratio is called the luminescence efficiency. The simplest picture of recombination can be obtained by assuming existence of electron traps below the edge of the conduction band and luminous centres above the edge of the valence band.

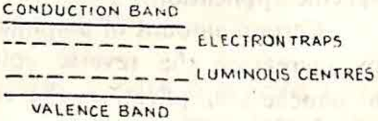


Fig. 7.14.

Let

n be the density of electrons in the conduction band

p—density of holes in the valence band

N_t—density of electron traps

N_l—density of luminous centres

n_t—density of electrons in electron traps

p_l—density of vacant energy states at the luminous centres

α_1—electron capture rate by traps

β—radiative recombination rate

γ_1—rate at which trapped electrons are released from electron traps by thermal agitation

δ – rate of non-radiative recombination of electrons from trap levels

If G is the generation rate we can write down the rate equations for the various energy levels.

Electrons in conduction band

$$0=\frac{dn}{dt}=G-\alpha_1(N_t-n_t)-\beta np_l+\gamma_1 n_t \tag{7.28}$$

For electrons in trap levels

$$0=\frac{dn_t}{dt}=\alpha_1(N_t-n_t)-\delta n_t p-\gamma_1 n_t \tag{7.29}$$

Adding equations (7.28) and (7.29) and rearranging we have

$$G=\beta np_l+\delta n_t p \tag{7.30}$$

The above expression indicates that the radiative recombination is in direct competition with non-radiative recombination. The ratio of the radiative recombination rate to the generation rate

$$\eta=\frac{\beta np_l}{G}=\frac{1}{1+\dfrac{\delta n_t p}{\beta np_l}} \tag{7.31}$$

If we assume that the Fermi level is situated in between the luminous centres and trap levels we have

$$n_t\propto N_t\exp\frac{E_f-E_t}{kT}$$

$$p_l\propto N_l\exp\frac{E_f-E_l}{kT}$$

$$\frac{n_t}{p_l}\approx\frac{N_t}{N_l}\exp\frac{E_l-E_t}{kT} \tag{7.32}$$

where E_l and E_t are the energies at the luminous centres and trap levels. Substituting the value of n_t/p_l in expression (7.31) we find that the efficiency is given by the expression

$$\eta=\left[1+\frac{\delta pN_t}{\beta nN_l}\exp\left(-\frac{E_t-E_l}{kT}\right)\right]^{-1} \tag{7.33}$$

The efficiency depends on the ratio N_t/N_l and the exponential term. Higher efficiencies can be achieved by increasing the density of luminous centres. The exponential term indicates that the efficiency decreases with increase in temperature.

The luminescence in semiconductors is the basis of diode lamps.

7.3.2 GaAs *diode lamp (light-emitting diode)*

When a GaAs *PN* junction is forward biased radiative recombination occurs near the junction and light of energy equal to the band gap is emitted. As the band gap of GaAs is 1.4 eV the wavelength of the emitted radiation is 8855 Å.

As the refractive index of GaAs is above 3.4 the critical angle is less than 18° and only light incident at angles less than 18° transmitted, rest of the radiation is internally reflected and absorbed when the junction is plane as in Fig. 7.15. To obtain maximum efficiency the geometry of the junction is designed in such a way that maximum amount of light emerges. One of the geometries is a spherical one for which the efficiency is large.

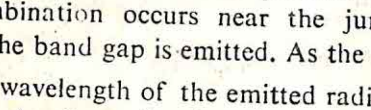


Fig. 7.15.

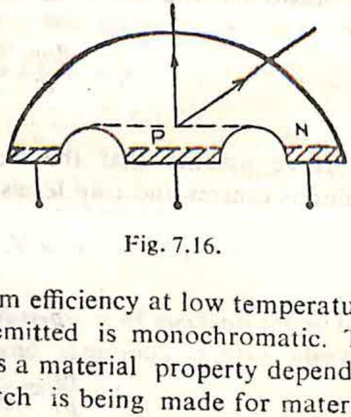


Fig. 7.16.

The diode lamps have maximum efficiency at low temperatures as stated earlier and the light emitted is monochromatic. The frequency of the emitted light is a material property depending on the band gap. Extensive search is being made for materials capable of emitting light at different frequencies. The response time for the diode is very short—of the order of nanoseconds, this property of the light-emitting diode makes it suitable for optical coupling in electronic circuits.

7.4.1 *Spontaneous emission—stimulated emission*

Light emission at a junction is due to electron transition from conduction band to vacant energy states. Different electrons in the conduction band fall to vacant energy states at different instants and the emitted photons are not time coherent. The emission is spontaneous. The number of transitions in unit volume of material in unit time or the rate of spontaneous emission is proportional to the density of electrons in the higher energy state.

Einstein proposed almost a quarter century earlier that the electrons in the higher energy state may be induced to fall into a lower energy state releasing photons of energy equal to the energy difference between the states. The emission is referred to as stimulated emission. One of the greatest advantages of stimulated emission is that all the emitted photons are time coherent as the electron transitions in stimulated emission occur at the same instant. The rate of stimulated emission is proportional to the product of the density of electrons in the higher energy state and the radiation density in the material.

As the stimulated emission is induced by radiation of the same frequency we have amplification. The amplification is called Light Amplification by Stimulated Energy Radiation or Laser.

7.4.2 *Two-level system*

Before proceeding to discuss Laser action at a junction we study in some detail the spontaneous and stimulated emissions in a two-level system.

Let E_1 and E_2 be two allowed energy levels between which electron transitions can occur and $E_1 - E_2 = h\nu$. Also let the densities of electrons in the two states be n_1 and n_2. The spontaneous emission rate$= A_{12} n_1$, where A_{12} is the probability of spontaneous emission.

If the material is located in radiation of density (u_ν) at the frequency ν, stimulated transitions occur and the stimulated transition rate is proportional to the product of the radiation density (u_ν) and the density of electrons in the higher energy state (n_1).

The stimulated emission rate$= B_{12} n_1 . (u_\nu)$, where B_{12} is the probability of stimulated emission.

As the material is located in radiation of density (u_ν) the radiation is absorbed by the material and electron transitions occur from the energy state E_2 to the higher energy state E_1. The absorption rate or the number of transitions from energy state E_2 to E_1 is proportional to the product of the radiation density (u_ν) and the density of electrons in the lower energy state (n_2). Therefore the absorption rate$=B_{21}\cdot n_2\,(u_\nu)$, where B_{21} is the probability of absorption. Under thermal equilibrium conditions the total emission rate equals the absorption rate and

$$A_{12}\,n_1+B_{12}n_1(u_\nu)=B_{21}\,n_2(u_\nu) \tag{7.34}$$

From the above equality the radiation density at the frequency is obtained as

$$u_\nu=\frac{A_{12}n_1}{B_{21}n_2-B_{12}n_1} \tag{7.35}$$

If the material is assumed to be in thermal equilibrium we have on application of Boltzman statistics.

$$n_1=n_2\exp\left(-\frac{E_1-E_2}{kT}\right) \tag{7.36}$$

From expressions (7.35) and (7.36) we find that

$$(u_\nu)=\frac{A_{12}}{B_{21}\exp\left(\dfrac{E_1-E_2}{kT}\right)-B_{12}} \tag{7.36}$$

The radiation density (u_ν) in the system is given by Planck's law as

$$(u_\nu)=\frac{8\pi h\nu^3}{c^3\left(\exp\dfrac{h\nu}{kT}-1\right)} \tag{7.37}$$

Expressions (7.36) and (7.37) agree only when $B_{12}=B_{21}$. Therefore it is obvious that the probability of absorption equals the probability of stimulated emission. Further

$$\frac{A_{12}}{B_{12}}=\frac{8\pi h\nu^3}{c^3} \tag{7.38}$$

The ratio is negligibly small at optical frequencies indicating that when a material is located in radiation of frequency ν absorption predominates spontaneous emission. If the thermal equilibrium is dropped, the number of transitions in unit time from the energy state E_1 to energy state E_2 is $A_{12}\,n_1+B_{12}\,(n_1-n_2)u_\nu$. The time rate of change in radiative density is given by the expression

$$\frac{du_\nu}{dt}=h\nu\,\{A_{12}n_1+B_{12}(n_1-n_2)u_\nu\} \tag{7.39}$$

The radiation density due to transitions from one energy state to another either decreases or increases with time depending on the factor $B_{12}(n_1-n_2)(u_\nu)$ as A_{12} is negligibly small. If $n_1 > n_2$ the radiation density increases with time according to the above expression. The radiation is stimulated and the emitted photons are all in phase.

Stimulated radiation can be obtained from a two level system when the density of electrons in the higher energy state is more than the density in the lower energy state (population inversion).

7.4.3 *Injection laser*

We studied in the last section the light emission in a forward biased GaAs *PN* junction. The emission is due to spontaneous transitions from the conduction band to the valence band, the radiation is spread over a frequency range and also non-coherent. Coherent radiation is obtained from a forward biased *PN* junction by providing an optical cavity at the junction.

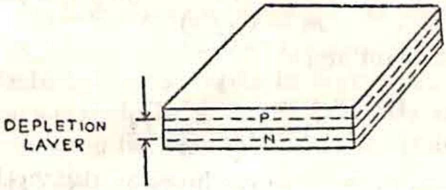


Fig. 7.17.

The opposite end faces of the junction are cleaved and polished so that the emitted radiation undergoes multiple reflections and the radiation density is maintained at an appropriate value. The material on either side of the junction is heavily doped making it degenerate. (The doping level is less than that in a tunnel diode) As the material on either side is degenerate the barrier voltage is more than the energy gap and it is possible to apply a forward voltage greater than the energy gap.

We now calculate the forward voltage needed to produce population inversion in the depletion layer where radiative recombination occurs.

The electron density in the conduction band within the depletion layer is given by the expression

$$n_c = C_1 f(c) \tag{7.40}$$

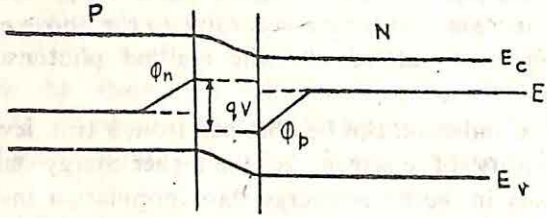


Fig. 7.18.

where C_1 is a constant and

$$f(c)=\frac{1}{1+\exp\dfrac{E_c-\varphi_n}{kT}} \tag{7.41}$$

φ_n being the quasi-Fermi level for the electrons on the P side. The electron density in the valence band within the depletion layer

$$n_2 = C_2 f(v) \tag{7.42}$$

where C_2 is a constant and

$$f(v)=\frac{1}{1+\exp\dfrac{E_v-\varphi_p}{kT}} \tag{7.43}$$

φ_p being the quasi-Fermi level for holes on the N side.

If (u_ν) is the radiation density at a frequency ν such that $h\nu > E_g$ the radiation is absorbed during electron transitions from the valence band to the conduction band. The number of transitions in unit time from the valence band to the conduction band is proportional to the product of the density of electrons in the conduction band, $C_2 \cdot f(v)$, density of vacant energy states in the valence band $C_1\{1-f(c)\}$ and the radiation density in the optical cavity (u_ν). The transition rate

$$\frac{dn_2}{dt}=W_{vc}\cdot C_1\cdot C_2\cdot f(v)\{1-f(c)\} \tag{7.44}$$

W_{vc} is the transition probability for transitions from the valence band to the conduction band.

Stimulated transitions occur from the conduction band to the valence band and the transition rate is proportional to the product of the density of electrons in the conduction band $C_1 f(c)$ density of vacant energy states in the valence band $C_2\{1-f(v)\}$ and the radiation density in the optical cavity (u_ν).

The transition rate

$$\frac{dn_1}{dt}=W_{cv}\cdot C_1\cdot C_2\cdot f(c)\{1-f(v)\} \tag{7.45}$$

W_{cv} is the transition probability for transitions from the conduction band to the valence band.

The radiative density in the cavity builds up when

$$\frac{dn_1}{dt} > \frac{dn_2}{dt}$$

As the transition probabilities W_{cv} and W_{vc} are equal we have from expressions (7.44) and (7.45) the condition for the buildup of radiation density as

$$f(c) > f(v) \tag{7.46}$$

Substituting the values of $f(c)$ and $f(v)$ from expressions (7.41) and (7.43) we have the condition for population inversion as

$$\begin{aligned} E_v - \varphi_p &> E_c - \varphi_n \\ \varphi_n - \varphi_p &> E_c - E_v \\ &> E_g \end{aligned} \tag{7.47}$$

$\varphi_n - \varphi_p$ is the applied forward voltage. We thus find that the voltage should be more than the energy gap. This is the reason why the semiconductor regions must be degenerate. The frequency of the radiation ν is such that $h\nu < E_g$. Often the stimulated radiation is due to transition between the band tails. (Band tailing occurs as the material on either side is degenerate.)

As the radiation of frequency ν propagates through the depletion layer in which there is population inversion the intensity of radiation increases due to stimulated transitions. If I_ν is the intensity of radiation at a frequency ν the radiation density is $u_\nu = \frac{I_\nu}{c}$ and the number of stimulated transitions in a distance dx is B_{12} $(n_1 - n_2)\ \frac{I_\nu}{c}\ dx$. The increase in energy due to stimulated transitions or the increase in energy flux

$$dI_\nu = h\nu B_{12}(n_1 - n_2)\frac{I_\nu}{c}\, dx \tag{7.48}$$

Substituting the value of B_{12} from expression (7.38) we have

$$dI_\nu = \frac{c^2}{8\pi\nu^2\mu^2} \cdot A_{12} \cdot (n_1 - n_2) I_\nu\, dx$$

As the wave is propagating in the semiconductor material the velocity of light is c/μ, where μ is the refractive index of the material.

Writing $A_{12} = \frac{1}{\tau_{\text{spont}}}$ (τ_{spont} being lifetime of spontaneous emission) and rearranging we have

$$\frac{dI/dx}{I} = \frac{c^2(n_1-n_2)}{8\pi\nu^2\mu^2\tau_{\text{spont}}} \tag{7.49}$$

The intensity of radiation increases as it propagates through the depletion layer. We can now define gain as the increase intensity for unit distance of propagation in a region of population inversion. The gain is given by the expression

$$g = \frac{c^2(n_1-n_2)}{8\pi\nu^2\mu^2\tau_{\text{spont}}} \tag{7.50}$$

At absolute zero of temperature the energy states above the quasi-Fermi level in the valence band are empty and $n_2 = 0$. n_1/τ_{spont} gives the number of radiative transitions occurring in unit volume of the active region (depletion layer). The gain expression can be written in a more useful form in terms of the current through the junction. If J is the current density through the junction, the number of electrons crossing unit area is J/q, and the number of photons emitted in the active region is $\eta J/q$, where η is the quantum efficiency. If d is the width of the depletion layer the number of photons emitted in unit volume is $\eta J/qd$. We can therefore write

$$\frac{n_1}{\tau_{\text{spont}}} = \frac{\eta J}{qd} \tag{7.51}$$

From expressions (7.50) and (7.51) we obtain the gain g as

$$g = \frac{\eta J c^2}{8\pi q\mu^2\nu^2 d}$$

We have been assuming till now that the emitted radiation is of a single frequency ν. This is not correct as the transitions do not occur between well-defined energy states. The radiation therefore has a definite band width $\Delta\nu$. If we take the band width into consideration the gain for unit frequency in the band is given by the expression

$$g = \frac{\eta c^2 J}{8\pi q\mu^2\nu^2 d\Delta\nu} \tag{7.52}$$

Having calculated the gain due to stimulated emission we now proceed to determine the condition under which coherent radiation is emitted from the junction. The radiation emitted at a point P in the active region travels towards the end where it undergoes reflection. After reflections at the

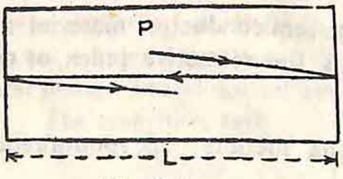


Fig. 7.19.

boundaries it reaches the point P again. During the round trip there is absorption in the material alongside gain. Also there is loss of energy at the reflecting ends. If I_0 is the intensity of radiation at the point P, α the absorption coefficient and R_1, R_2 the reflection coefficients of the end faces, the intensity of the radiation after one round trip in a cavity of length L is given by the expression

$$I = I_0 R_1 R_2 \exp[2(g-\alpha)]L \qquad (7.53)$$

Coherent radiation is obtained from the cavity when $I = I_0$ or when

$$1 = R_1 R_2 \exp[2(g-\alpha)]L \qquad (7.54)$$

writing $R = \sqrt{R_1 R_2}$.

The condition for emission of coherent radiation from the cavity is

$$J = \alpha + \frac{1}{L} \ln \frac{1}{R} \qquad (7.55)$$

Substituting the value of g from expression (7.52) we find threshold current density necessary for lasing

$$g_{th} = \frac{8\pi q \mu^2 \nu^2 \Delta\nu}{\eta c^2} \left\{ \alpha + \frac{1}{L} \ln \frac{1}{R} \right\} \qquad (7.56)$$

$$= \frac{1}{\beta} \left\{ \alpha + \frac{1}{L} \ln \frac{1}{R} \right\}$$

where $\beta = \dfrac{\eta c^2}{8\pi q \mu^2 \nu^2 \Delta\nu}$ (7.57)

The above equations are valid at 0°K only as it has been assumed earlier that energy states above the quasi-Fermi level are vacant. However the expression gives an idea of the threshold current density necessary for lasing.

.The current density is about 80 amp/cm^2 at liquid helium temperatures but several hundreds of times at the room temperature. This is because the absorption coefficient increases rapidly with temperature. Junction lasers are continuously operated at liquid helium and liquid nitrogen temperatures. Continuous operation at room temperature involves large threshold currents and considerable heat dissipation at the junction. At room temperatures the lasers are operated only in pulsed conditions. For current densities less than the threshold current density the absorption in the cavity is more than the gain and the light emitted is due to spontaneous emission, as such the band width of the emitted radiation is wide. When the current density reaches the threshold value the gain equals the losses in the cavity. As the gain is inversely proportional to the square of the frequency the stimulated emission is restricted to a very narrow band about the central frequency. The radiation is monochromatic.

Population inversion in a semiconductor can also be achieved by optical pumping or by electron bombardment.

We have discussed GaAs injection laser in some detail. At present a number of semiconducting materials, which can lase at different frequencies, are available. The requirement for a material suitable for lasing is that it should be a direct band gap semiconductor and also capable of being doped to form a junction. Extensive search is being made for materials that can be used for lasing.

Lasers are used in the fields of communication, holography, medicine etc., and the potentialities of its use are vast. It is not possible in an introductory book on semiconductor devices to discuss the research and technological applications. An attempt is made in this book to present the basic principle of operation of an injection laser.

Important Formulae

1. Change in conductivity of a sample exposed to light

$$= d\sigma = q\eta\varphi(\tau_p\mu_p + \tau_n\mu_n)$$

2. Gain of a photoconductor $G = \frac{\tau\mu V}{L^2}$

Average power output of a photoconductor

$$= \frac{1}{8}(\eta q\varphi_1 A)^2 \frac{1}{G(\omega t)^2}$$

3. Current voltage relation of an illuminated junction

$$I = I_0\left(\exp\frac{qV}{kT} - 1\right) - I_L$$

4. Efficiency of a solar cell

$$= \frac{I_0 q I_L V^2_{\text{mp}}}{AkT\left(1 + \frac{qV}{kT}V_{\text{mp}}\right)}\left(1 + \frac{I_0}{I_L}\right)\frac{1}{P_{\text{in}}}$$

5. Maximum power available from a photodiode

$$= \left| q\eta A\varphi(\omega) \right|^2 \frac{1}{\omega^2 c^2 R}$$

6. Luminescence efficiency

$$= \left[1 + \frac{\delta p N_t}{\beta n N_t} \exp\left(-\frac{E_t - E_i}{kT}\right)\right]^{-1}$$

The ratio of the spontaneous to stimulated emission probabilities

$$\frac{A_{12}}{B_{12}} = \frac{8\pi h\nu^3}{c^3}$$

Gain in the optical cavity of an injection laser

$$g = \frac{\eta c^2 J}{8\pi q \mu^2 \nu^2 d \Delta\nu}$$

Threshold current density in an injection laser

$$J_{th} = \frac{8\pi q \mu^2 \nu^2 \Delta\nu}{\eta c^2}\left[\alpha + \frac{1}{L} \ln \frac{1}{R}\right]$$

Questions

1. What is the necessary condition for optical generation of charge carriers? How is this phenomenon used to determine the band gap of a semiconductor?
2. Derive an expression for the change in conductivity of a sample of intrinsic semiconductor when exposed to light of intensity φ photons/cm². sec.
3. What is meant by the gain of a photoconductor? Derive an expression for the gain in terms of the length of the sample and the voltage across it.
4. Find an expression for the average power output from a photoconductor when used to detect intensity modulated optical signals.
5. Derive the current voltage relation for an illuminated junction. From the expression evaluate: the open circuit voltage; short circuit current.
6. Explain the condition under which an illuminated junction delivers power.
7. Derive an expression for the efficiency of a solar cell.
8. Draw the equivalent circuit of solar cell. How is the efficiency affected by the series resistance?
9. Explain the operation of a photodiode.
10. Draw the equivalent circuit of a photodiode when used to detect intensity modulated signals and evaluate the maximum power available.
11. Compare the response time of a photodiode with that of a photoconductor.

12. What is meant by electroluminescence? What is the requirement to be satisfied for luminescence in a semiconductor?
13. Show that non-radiative recombination is in direct competition with radiative recombination.
14 What is meant by luminescence efficiency? Derive an expression for the luminescence efficiency of a semiconductor.
15. Describe the construction of GaAs diode lamp. Where is it used?
16. Distinguish between spontaneous and stimulated emission.
17. Evaluate the ratio of the probabilities of spontaneous and stimulated emission.
18. What is meant by population inversion?
19. What is the necessity of optical cavity in a laser?
20. For what forward voltage does population inversion occur in a forward biased junction?
21. Derive an expression for the gain in the optical cavity of an injection laser.
22. Evaluate the threshold current density at which lasing occurs in a junction laser.

Index

Admittance of a junction 48
Auger process 17
Avalanche breakdown 45
  multiplication 63, 165
  region 170

Band tailing 176
Barrier voltage 27
Base spreading resistance 64
  width modulation factor 113
Bloch function 2
Brillouin zone 3

Capacitance junction 51
Capture cross section 17
Carrier generation 15
  recombination 15
Channel 120
Charge storage 48
Common emitter current gain 68
Conduction band 4
Contact potential 29
Continuity equation 21
Crystal momentum 2
Cut in voltage 44
Cut off frequency α 100
  β 94, 100
  dynamic 159

Dark resistance 197
Decay time 53, 107
Degenerate amplifier 157
Depletion layer 28, 30
  mode 144
Delay time 106
Dielectric relaxation
  time 183
Diffusion 14
  constant 15
  length 22, 23
Doped semiconductor 8
Drain 120, 141
Drift 11
  region 173
  transistor 108

Early effect 81
Ebers-Moll equations 101
Effective mass 6
Einstein's equation 15
Electron trap 17
Emitter efficiency 58, 59
Enhancement mode 144
Extrinsic semiconductor 8

Fermi Dirac function 7
  level 7, 27
Field effect transistor 119
Flat band voltage 137
Forbidden gap 4

Gain-band width product 94
Gate 120, 141
Graded junction 33
Gunn effect 180, 184

Hole 4
Hot electron 12
Hyper abrupt junction 159

IMPATT 169
Impurity scattering 13
Imrefs 40
Inversion 134
Ionisation integral 47
  rate 46

Laser 209
Lattice scattering 13
Life time 20
Light detecting resistors 194
Luminescence 205

Manley Rowe equations 156
Memory element 149
Mobility 12
Multiplication factor 47
Mutual conductance 91

Non-radiative recombination 17

Optical generation 15, 193

Parametric amplification 153
Photoconductivity 193
Photoconductor 194
Photodiode 197, 203
Photo voltaic effect 197
Pinch off 121, 144
PIN diode 161
Punch through 72, 114

Quantum efficiency 194
Quasi-Fermi level 40

Radiative life time 17
  recombination 16
Read diode 169
Recombination centre 17
Response time 127, 197
Rise time 107
RWH mechanism 180

Schrodinger's equation 1
Semiconductor direct band gap 5
Semiconductors 4
Shockley equation 36
  and Read theory 17
Snap back diode 160
Solar cell 200
Source 120, 141
Spontaneous emission 208
Square law device 126, 147
Stimulated emission 208
Storage time 53, 107
Stored base charge 74
  charge 51
Surface recombination 20
    velocity 21
Sustaining voltage 73

Thermal generation 15
  resistance 96
Transconductance 124, 148
Transport factor 59, 62
Transit time 84
Turn off voltage 123
  on time 106
  on voltage 141
Tunnel diode 175

Up convertor 157

Valence band 4
Varactor 158

Wave vector 2

Zener breakdown 45

S. C. E. R. T. W
Library

Introduction to Electronics

K.J.M. Rao

This is a textbook of electronics with a difference. It caters to the needs of the average Indian undergraduate student.

Such rapid developments are taking place in this branch of physics that the classical treatment of vacuum tube circuits is quite inadequate and often unrewarding. The present treatment starts with the fundamentals and develops the subject logically and systematically to afford the reader a glimpse of the latest developments that are taking place. An integrated approach to electronics, both vacuum tube and solid state, is maintained throughout. A lot of deadwood has been removed to make it easy for the student to understand the important principles and thus make quicker progress towards the modern applications. Simple mathematical treatment is given to afford quantitative confirmation of basic principles.

Other Interesting Books

ELECTRONIC DEVICES AND THEIR APPLICATIONS

P.N. Das

The most authoritative book available on the subject playing a vital role in all the fields of electrical engineering and for investigation work in any branch, dealing with all different types of electronic devices. It covers all fundamental aspects of electronics starting from principles of ballistic electronics, includes all the major applications to make the book suitable for the beginners to start with and to give comprehensive knowledge on the subject and its applications. As transistors are now widely used in all fields, their applications are given side by side with the applications of vacuum tube wherever possible.

TEXTBOOK OF ELECTRICITY AND MAGNETISM (12/E)

D.P. Acharya

An important and well established textbook exactly suitable for a student preparing for the degree course in an Indian University. In making this book exactly suitable for a student preparing for the degree course in an Indian University, the author has deliberately cut down all unimportant elementary portions which are usually done in the intermediate stage. In this edition the book has been thoroughly revised and the latest questions have been added at the end of each chapter.

OXFORD & IBH PUBLISHING CO. PVT. LTD.
NEW DELHI BOMBAY CALCUTTA